Aalto University
School of Electrical Engineering
Degree Programme in Bioinformation Technology

Juhani Kähärä

Using DNase I hypersensitivity Data for Transcription Factor Binding Predictions

Master's Thesis
Espoo, May 24, 2014

Supervisor: Harri Lähdesmäki
Instructor: Harri Lähdesmäki

Aalto University
School of Electrical Engineering
Degree Programme in Bioinformation Technology

ABSTRACT OF
MASTER'S THESIS

Author: Juhani Kähärä

Title:
Using DNase I hypersensitivity Data for Transcription Factor Binding Predictions

Date:	May 24, 2014	**Pages:**	vii + 58
Professorship:	Information and Computer Science	**Code:**	T-61
Supervisor:	Harri Lähdesmäki		
Instructor:	Harri Lähdesmäki		

Transcription is a key information process in the cell and transcriptional regulation is largely controlled by DNA binding proteins called transcription factors. Understanding transcription factor binding is integral to understanding the most important biological events, such as gene expression and the function of gene regulatory networks.

Currently transcription factor binding sites are determined by chromatin immunoprecipitation followed by sequencing, but this method has several limitations. To overcome these caveats, DNase I hypersensitive sites sequencing is increasingly being used for mapping gene regulatory sites. Computational tools are needed to accurately determine transcription factor binding sites from this new type of data.

In this work a novel method, BinDNase, is developed for detecting transcription factor binding sites using DNase I hypersensitivity data. The method is applied to 57 different transcription factors in cell type K562. We demonstrate that the prediction performance of BinDNase exceeds the performance of other existing methods.

Our results indicate that DNase I hypersensitivity data should be used in multiple resolutions instead of the highest possible resolution. We also show that the binding predictions should be made separately for each transcription factor and that the sequencing depth of currently available data sets is sufficient for binding predictions for most transcription factors. Finally, we show that models built with BinDNase generalize between different cell types making the method a powerful tool in transcription factor binding predictions using DNase I hypersensitivity data.

Keywords:	transcription factor, binding modeling, DNase-seq, DNase I hypersensitivity, DNase footprints
Language:	English

Aalto-yliopisto
Sähkötekniikan korkeakoulu
Bioinformaatioteknologian tutkinto-ohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Juhani Kähärä

Työn nimi:
DNase I hypersensitiivisyysdatan käyttö transkriptiotekijöiden sitoutumisennusteissa

Päiväys:	23. toukokuuta 2014	**Sivumäärä:**	vii + 58
Professuuri:	Tietojenkäsittelytiede	**Koodi:**	T-61
Valvoja:	Harri Lähdesmäki		
Ohjaaja:	Harri Lähdesmäki		

Transkriptio on solujen välttämätön informaatioprosessi ja transkriptiota säädellään pääasiassa DNA:han sitoutuvilla proteiineilla, joita kutsutaan transkriptiotekijöiksi. Transkription ymmärtäminen on elintärkeää ymmärtääksemme tärkeimpiä biologisia toimintoja, kuten geenien säätelyverkostojen toimintaa.

Nykyään transkriptiotekijöiden sitoutumiskohdat määritetään sekvensoimalla geneettinen materiaali kromatiinin vasta-ainesaostuskokeesta, mutta tällä menetelmällä on useita heikkouksia. Näiden ongelmien vuoksi DNaasi I hypersensitiivisten alueiden sekvensointia käytetään enenemässä määrin geenien säätelyalueita etsittäessä. Laskennallisia menetelmiä tarvitaan määrittäämään transkriptiotekijöiden sitoutumiskohdat tarkasti käyttäen tätä uudenlaista dataa.

Tässä työssä kehitettiin uusi menetelmä, BinDNase, transkriptiotekijöiden sitoutumisennusteiden tekoon käyttäen DNase I hypersensitiivisyysdataa. Menetelmää käytettiin ennustusten laatimiseen 57 eri transkriptiotekijälle solutyypissä K562 ja BinDNase:n ennusteet ovat tarkempia kuin muiden olemassa olevien menetelmien.

BinDNase:lla saadut tulokset viittaavat siihen, että DNase I dataa pitäisi käyttää usealla eri resoluutiolla tarkimman mahdollisen resoluution sijaan. Tässä työssä osoitetaan, että ennusteet pitäisi tehdä erikseen kaikille transkriptiotekijöille ja että sekvensointisyvyys jo olemassa olevissa aineistoissa on riittävä ennustusten tekoon useimmilla transkriptiotekijöillä. Näytämme vielä, että BinDNase:lla rakennetut mallit yleistyvät toisille solutyypeille. Tämä tekee menetelmästä tehokkaan työkalun transkriptiotekijöiden sitoutumisennusteiden tekoon käyttäen DNase I hypersensitiivisyysdataa.

Asiasanat:	transkriptiotekijä, sitoutuminen, sitoutumisennusteet, DNase-seq, DNase I hypersensitiivisyys
Kieli:	Englanti

Acknowledgements

I wish to express my deepest gratitude towards all who have contributed positively to the completion of this master's thesis. First of all, I would like to thank my supervisor and instructor professor Harri Lähdesmäki for the immense amount of ideas, advice and support required in this work. Professor Lähdesmäki has given lots of encouragement and inspiration throughout the entire thesis project.

I would like to also thank the whole Aalto Systems biology group for great working atmosphere and the administrators of the Triton computing cluster in the Aalto department of Information and Computer Science for providing proper resources for computational work.

I wish to thank all my relatives and friends for supporting me during the thesis project and throughout my life. Special thanks are awarded to Joona Jäntti for company on relaxing coffee breaks and inspirational talks about his impactful work on cognitive radios.

Finally, I would like to thank Leevi Letkutsalo, the hose master of Retuperän WBK, for keeping Otaniemi and the whole Espoo area safe from flammable and other dangerous substances.

Espoo, May 24, 2014

Juhani Kähärä

Abbreviations and Acronyms

AUC	Area under the ROC curve
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
DNase I	Deoxyribonuclease I
DNase I HS	DNase I hypersensitivity
FPR	False positive rate
HLH	Helix-loop-helix
PBM	Protein binding microarray
PSFM	Position specific frequency matrix
PWM	Position weight matrix
ROC	Receiver operating characteristic
SELEX	Systematic evolution of ligands by exponential enrichment
TF	Transcription factor
TPR	True positive rate

Contents

Chapter 1

Introduction

Life consists of self-replicating machines made out of chemical substances. A typical living machine is the cell, the basic building block of all animals, plants and fungi. All living creatures require an information processing system for replicating. The information processing system is often incorporated to facilitate the replicating process with functions such as metabolism, movement and adaptation to different environmental conditions.

The central information processing system in modern life is presented in Figure 1.1. The figure describes the information flow from DNA to protein through RNA intermediate. The key steps in this central dogma of biology are transcription from DNA to RNA and translation from RNA to proteins. These steps are highly sophisticated and numerous modification and regulation measures can be used to alter these processes. This work studies one of the main mechanisms affecting transcription.

Transcriptional regulation is largely controlled by transcription factors (TFs) that bind short (10-20 bp) DNA sequence motifs in gene promoters, enhancers and other regulatory sites. Many TFs bind DNA in a sequence specific manner and understanding TF binding is integral to understanding gene regulatory networks. Moreover, changes in the genomic DNA at TF-DNA interaction sites can affect TF binding and can contribute to phenotypic differences, including gene expression [13], but can also contribute to various diseases [21, 34]. Determining the locations of TF-binding sites is therefore of high importance.

The current state-of-the-art method for genome-wide profiling of TF-binding is chromatin immunoprecipitation followed by sequencing (ChIP-seq). However, ChIP-seq has some short falls, such as it is possible to map the positions of only one TF per experiment. The ChIP-seq protocol is discussed in more detail in Section 2.3.

Another protocol called DNase I hypersensitivity experiment followed by

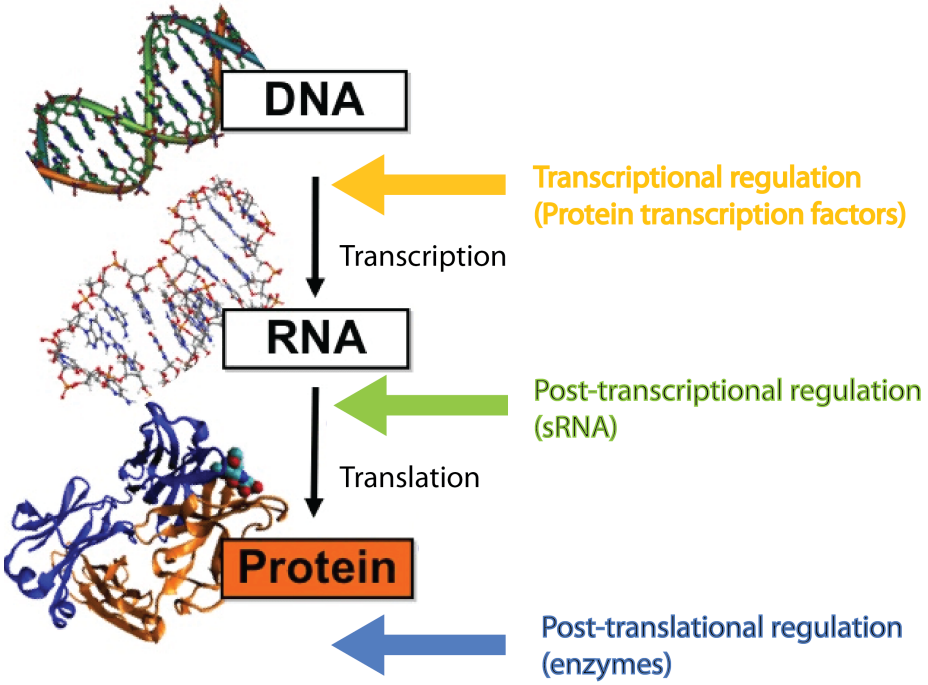


Figure 1.1: The central dogma of modern biology describes the information flow from DNA to RNA and from RNA to proteins. This process can be regulated in all stages. (The figure is taken from website in [1])

sequencing (DNase-seq) detects a signal at open chromatin sites genome-wide. Consequently, DNase-seq is increasingly used to complement ChIP-seq experiments because a single DNase-seq experiment can provide valuable information about putative TF-DNA interaction sites for all TFs. Genome-wide maps of putative regulatory sites in selected cell types/lines detected using DNase-seq data have already been created e.g. in the ENCODE project [22]. DNase-seq has the potential to replace ChIP-seq in genome wide TF-binding site profiling. DNase I hypersensitivity is discussed in Chapter 3.

Currently, the exact locations of TF-binding events are pinpointed by finding stereotypic DNase I footprints. These footprints are short genomic locations of low DNase I cleavage activity immediately flanked by high DNase activity. An illustration of such regulatory site is shown in Figure 1.2 a) where the ATF1 motif locations within ChIP-seq peaks are characterised with low DNase activity and the flanking regions exhibit high DNase activity. It has however been reported that for some proteins nucleotides in the middle of TF-DNA interface are exposed to DNase I cleavage [22]. Therefore, treating

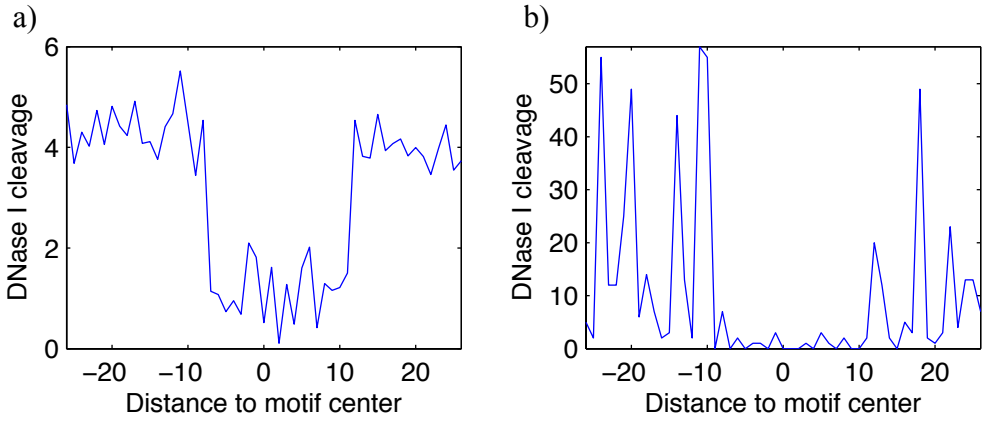


Figure 1.2: a) The average DNase I cleavage around ATF1 binding sites resembles the canonical definition of DNase I footprint. The average DNase-seq signal at nucleotide resolution centered at ATF1 motif overlapping ATF1 ChIP-seq peaks is shown. b) DNase-seq signal at nucleotide resolution around a single AFT1 binding site located between nucleotides 96,929,096–96,929,148 in chromosome 9 in celltype K562. Although individual footprints are noisy, the canonical shape of the footprint is still visible in the data.

all the nucleotides as protected in the TF-DNA interface might not be an adequate way to model the binding.

The DNase I footprints have been shown to differ between TFs and they are speculated to contain nucleotide resolution information. However, most of the methods developed for identifying footprints use the canonical definition of DNase footprints of low DNase activity flanked by high activity [18, 22, 24], although some methods include nucleotide resolution information [25].

On the other hand, a recent paper shows that the nucleotide resolution DNase I cleavage pattern is partly caused by the intrinsic sequence bias of the DNase molecule [7] suggesting that the nucleotide resolution DNase-seq signal at the TF-DNA interaction site do not necessarily provide predictive power to distinguish real binding sites. The sequence bias of the DNase I experiment is discussed in Section 3.2. Moreover, the DNase I footprint signal at individual genomic locations is noisy as illustrated in Figure 1.2 b) which shows the DNase signal in one genomic location. Consequently, carefully designed computational methods are needed for DNase-seq data processing.

Having the aforementioned advantages and disadvantages of DNase-seq data in mind, here we study the use of high resolution DNase I hypersensitivity data for predicting TF binding sites. We develop a method which,

for each TF, automatically extracts features from the DNase-seq data which maximally discriminates bound and unound genomic locations. The method will be compared with previously published methods designed for this task and many characteristic features of the DNase I hypersensitivity data, such as the required sequencing depth and the best resolution to be used in the modeling, are investigated. The results obtained with the developed method, BinDNase, shed light on many questions considering the use of DNase I hypersensitivity data in TF-binding modeling. The results are discussed in detail in Chapter 6.

A full length article describing the method and key findings presented in this work was submitted to the ECCB 2014 conference.

Chapter 2

Sequence specific transcription factors

2.1 The role of transcription factors in the cell

Transcription factors (TFs) are DNA binding proteins that control transcription. Sometimes they are called sequence specific transcription factors for the fact that they bind DNA in a sequence specific way. There is a great variety of TF function and structure as 2600 human proteins contain DNA binding domains [4]. All of these DNA binding domain containing proteins are not necessarily transcription factors and an article from 2009 reports 1391 high confidence TFs [29]. The true number of TFs is most likely between these two numbers.

Transcription factors have a significant role in key biological functions such as development, cellular processes and stimulatory response [29]. The wide range of cell type specific transcription factors are essential for cell differentiation and proliferation. TFs act by binding a regulatory region in the genome which leads to either promoting or repressing the transcriptional activity of their target gene(s). This is illustrated in Figure 2.1. Some of the main mechanisms which TFs use for transcriptional regulation include:

- Attract or repel the transcriptional machinery
- Regulation of chromatin state
- Prevent other TFs from binding
- Attract other TFs to bind

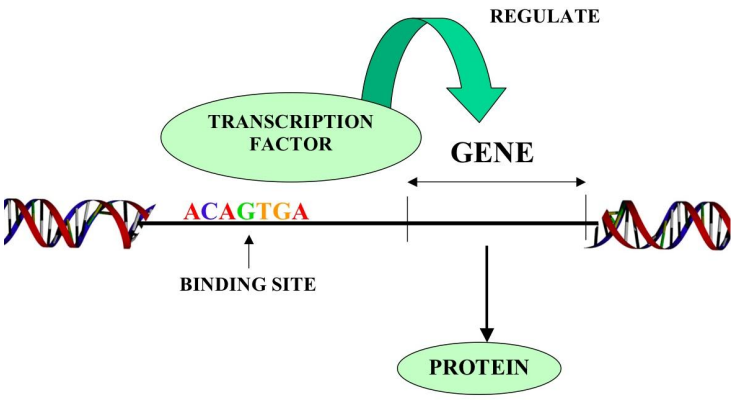


Figure 2.1: Transcription factors regulate gene expression by binding to regulatory sites. The picture is adopted from the website in [2].

The sequence specificity of each TF is determined by its 3D-structure and the DNA-binding domain of the TF. The DNA-binding domain is used to distinguish a specific sequence in the DNA. Typically these bound signal sequences are found in the major groove of the DNA, where the nucleotides are exposed for TF recognition [23]. Because TF binding is a form of structural recognition, the TFs are most commonly classified according to their structure. One such classification of TFs identifies nine different structural super classes [33]. The names of these classes and the proportional sizes of each class can be seen in Figure 2.2. All the superclasses exhibit distinguishable structural similarity in the way they bind to the DNA. Each class can further be divided into smaller groups. In many cases structural similarity in TFs indicates similar DNA sequence specificity. The TF structure might also be a factor in choosing the best way to model the TF-DNA interaction.

Variation in transcription factor binding can lead to phenotypic differences [13]. The variation is caused by mutations that directly disrupt the sequence that the TF would bind [13, 16]. Other mechanisms for differential TF binding, such as the variation in chromatin state modulation, are likely to exist as well, but this question is not yet researched in detail. The phenotypical changes are highly expected as transcription is one of the key biological processes. Some mutations affecting TF binding has also been linked to various diseases [16, 21, 34]. Due to the important role of TFs in the cell, it is highly important to be able to predict the binding sites for each TF. Mapping TF binding sites accurately leads to better understanding of basic biology and has the potential to benefit human health by revealing disease related differences in TF binding.

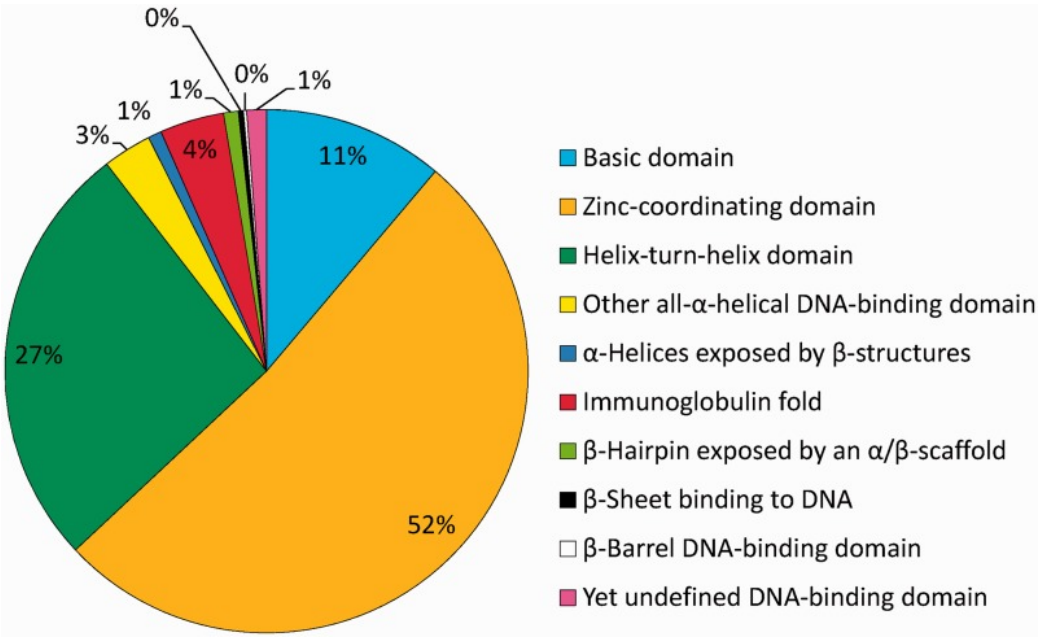


Figure 2.2: The TF super classes and their proportional sizes [29].

2.2 Transcription factor binding modelling

The sequence specifity of transcription factors can be modelled in numerous ways. A publication from 2013 evaluates 26 different methods for making TF binding predictions [31]. The models can be divided in different groups based on their main characteristics. Three of such model groups are position specific matrix models, k-mer models (for example [3, 11]) and dinucleotide models.

All the different models require experimental information of the binding specificity of the protein of interest. This information can be learned using different experimental protocols such as protein binding microarrays (PBM) [5], chromatin immunoprecipitation followed by sequencing (ChIP-seq) [20] or systematic evolution of ligands by exponential enrichment (SELEX) [9, 10]. It is under speculation which model works the best for each application and what kind of experiments are required to build the models [11, 31]

The most common way to model the sequence specificity is the position specific frequency matrix (PSFM). The frequency matrix lists the probability of each nucleotide in each position in the binding site. Each position is considered independent from other positions. This independency assumption is known to be false because there are reported dependencies between nucleotides at least for some TFs [10]. Dinucleotide models and k-mer

models try to take these dependencies into account but nevertheless these model classes do not show better prediction performance in all cases. This leaves PSFM modeling the state-of-the-art method for basic computational TF binding prediction model.

The binding modeling that utilizes PSFMs is called scanning (also motif scanning, PSFM scanning etc.). The matrix model of lenght l is laid on top a sequence of interest and the matrix is used to calculate the binding score. The binding score reflects the presumed affinity of a TF to a sequence. The mathematical details for the calculating binding scores will be explained in Section 5.1. If the sequence of interest is longer than l, the matrix model is slid to the next position and the binding strength is evaluated again. Typically, motif scanning for larger genomic regions yields extensive numbers of false positive binding sites because any sequence can be found in larger regions purely by chance. Additionally, only certain binding sites are accessible for TFs due to chromatin constraints. For these reasons the binding modeling has to use additional data to narrow down the possible binding sites to produce meaningful knowledge about TF binding. Chapter 3 will shed light on how DNase I hypersensitivity is used as additional information in binding modeling.

2.3 ChIP-seq

The current state-of-the-art method for mapping transcription factor binding sites in a whole genome scale is chromatin immunoprecipitation followed by sequencing (ChIP-seq). The method can map TF binding sites genome wide in a single experiment. It has been used in hundreds/thousands of publications and it is the main TF binding mapping method in the ENCODE consortium [28].

The workflow of ChIP-seq is presented in Figure 2.3. The experiment begins by crosslinking all DNA-bound proteins to the DNA with formaldehyde. In the next step, the DNA is isolated from the samples and the DNA is sonicated to produce DNA fragments. Some of these fragments are bound by DNA binding proteins. DNA-protein complexes of interest are immunoprecipitated using a specific antibody. The precipitated complexes are then purified, the DNA-protein crosslinks are reversed and the remaining DNA sequences are sequenced using a high-throughput sequencer.

The standard in high-throughput ChIP-seq experiments is to produce at least 10 million uniquely mapping sequencing reads [14]. The reads have to be processed in a computational workflow to obtain the TF binding sites. Typically the resulting binding regions from ChIP-seq experiments span several

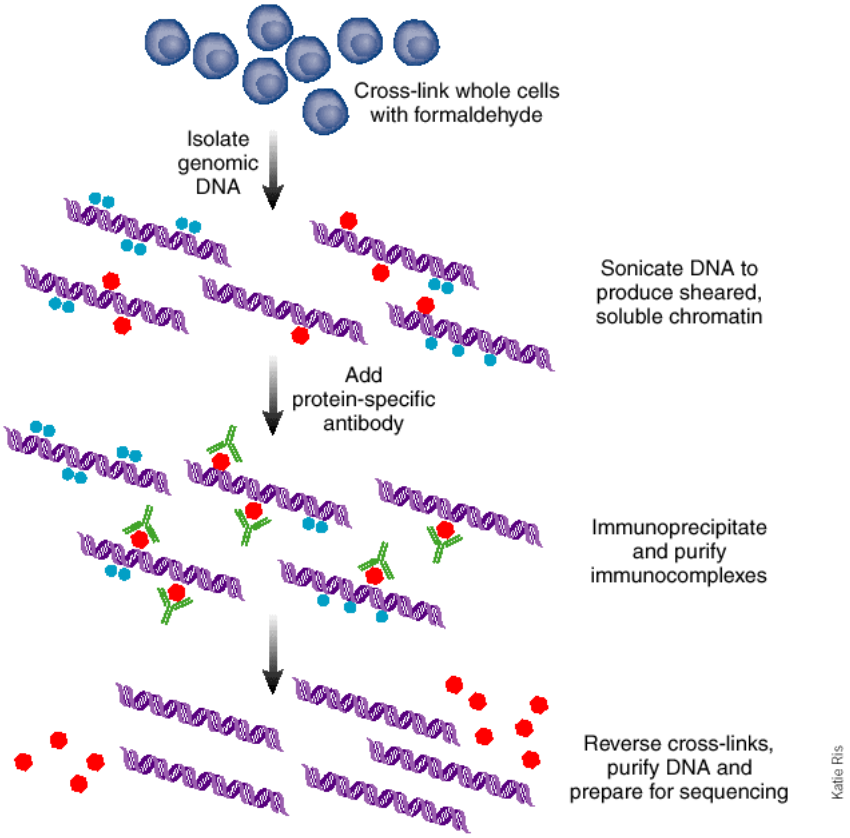


Figure 2.3: ChIP-seq workflow [20]

hundred nucleotides and therefore cannot pinpoint the exact location of the TF-DNA interaction. Computational approaches such as PSFM-modeling can be used to scan the ChIP-seq regions to accurately determine the binding site, although this method does not always yield useful results.

The ChIP-seq experiments would be time consuming and expensive if we wanted to know the binding locations of many or all transcription factors binding to the human genome. Furthermore, the experiment requires expensive antibodies and some TFs could lack such ChIP-grade antibody completely. Additionally, several antibodies may exist for a single TF and the quality of the antibodies vary heavily. The low resolution of ChIP-seq is also a serious issue if we would like to build an accurate map of TF binding events.

The main shortfalls of ChIP-seq are:

- Mapping only one factor per experiment
- Low resolution
- Requirement of a ChIP-grade antibody

To overcome these caveats, other methods are being developed to either complement or replace ChIP-seq in mapping TF binding locations. One of such contemporary protocols is DNase-seq, which addresses all of these shortfalls and for this reason the method is studied in this work. DNase-seq is a method for detecting DNase I hypersensitivity on a whole genome scale. The protocol and some of its characteristics are discussed in the following Chapter 3.

Chapter 3

DNase I hypersensitivity

3.1 DNase I activity indicates open chromatin

Deoxyribonuclease I (DNase I) is a protein that has the ability to cut DNA. Its main role in the cell is speculated to be waste management and DNA disposal during apoptosis [26], but the molecule is increasingly used in biotechnical and research applications. A widely used application is detecting DNase I hypersensitive sites which are often considered as regulatory regions.

DNase I hypersensitive sites sequencing (DNase-Seq) is an experimental protocol for determining which genomic regions are available for DNase I digestion. In the experiment a suitable concentration of the DNase I molecule is added to the samples which leads to releasing DNA fragments of different lengths. The DNA fragments are size-selected by taking the short sequences ($<$500 bp long in [22]) or by practising more rigorous size selection by dividing the short sequences further in different categories (50-100 bp, 100-200 bp and 200-300 bp long in [7]). The size-selected fragments are then subjected to high-throughput sequencing and the sequenced reads can be analyzed with computational methods.

The rationale for the experiment is that the active regulatory sites are located in open chromatin regions. On the other hand the inactive form of DNA, heterochromatin, is tightly wrapped around histones and is therefore protected from DNase I cleavage. The DNase molecule cannot reach a suitable position for cutting the DNA if the site is occupied by histones or other DNA-binding molecules. For this reason the DNase I experiment gives information about openness of the chromatin.

It is also observed that the exact locations where transcription factors bind are protected from DNase I. This has lead to the identification of DNase I footprints as the pinpointed transcription factor binding sites. The foot-

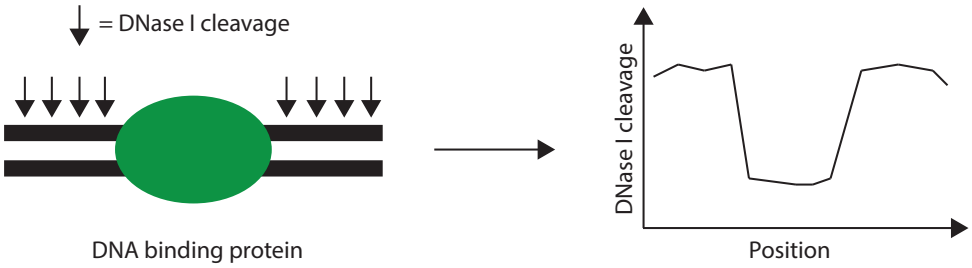


Figure 3.1: In this simplified view of TF-DNA interaction the protein is tightly bound to the DNA and protects the underlying sequence from DNase I digestion. The binding event would produce a traditional TF footprint in the DNase I cleavage signal.

print is a relatively short genomic site (8-30 bp) which is characterized with low DNase activity but which is immediately flanked by high DNase activity. In this work, this basic definition of DNase I footprint is referred to as the canonical or traditional definition of a footprint. Most of the existing methods identifying TF binding sites using DNase-seq data rely on this definition [7, 18, 19, 22, 24]. A simplified model of TF-DNA interaction that results in the canonical TF footprint is shown in Figure 3.1.

In a more realistic view, the underlying DNA is not completely protected from the DNase cleavage and finer details of TF-DNA interactions can readily be seen in deeply sequenced DNase I libraries. An article from 2012 [22] reports one nucleotide resolution patterns at regulatory sites that could be resulting from the binding of specific TFs to these sites. This is illustrated in Figure 3.2. The binding of protein USF1 to DNA results in a traditional footprint pattern in which there are two distinct DNase cleavage peaks right in the middle of the binding event. According to the authors, these two peaks are caused by nucleotides in the TF-DNA interface that are exposed to DNase digestion. Many similar patterns can be seen in the data ([22], Section 4.3, Appendix B).

DNase-seq can be applied to finding regulatory sites and even specific TF binding sites by using computational methods. One way to assign a specific TF to a regulatory site found by DNase-seq experiment is to scan the regions with known binding motifs. There are also methods that predict TF binding by looking directly at the DNase I activity within the potential binding sites [18, 24, 25]. By further developing the biological protocol and computational tools DNase-seq has the potential to replace ChIP-seq as the state of the art method for mapping TF-DNA interactions. The main advantages of DNase-seq over ChIP-seq are:

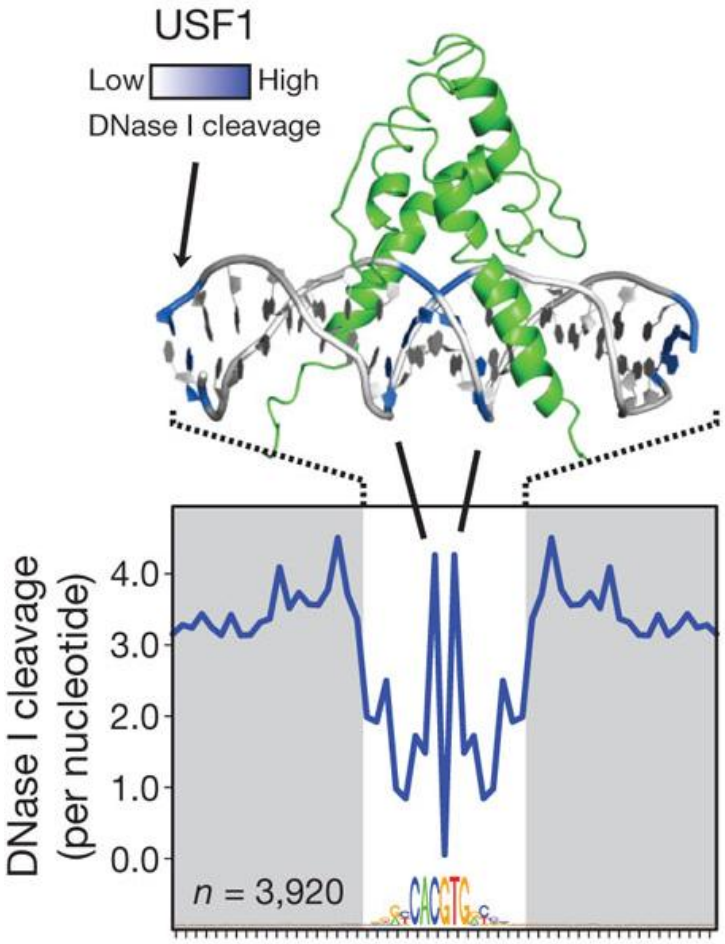


Figure 3.2: The average DNase I cleavage at the binding sites of USF1. The average DNase I cleavage at USF1 binding sites resemble the canonical definition of the TF footprint, but there are two high peaks in the middle of the footprint. These peaks do not fit to the traditional footprint definition. The USF1-DNA X-ray structure is also shown in the figure. The two peaks right in the middle of TF-DNA interface might be result of exposed nucleotides as seen in the structural figure. [22]

- Potential to map the binding sites of all TFs
- Single nucleotide resolution
- No need for high quality antibodies

3.2 DNase I sequence bias

A study from 2014 reports that the highly stereotypic cleavage patterns might be resulting from the intrinsic sequence bias of the DNase I molecule instead of DNA-protein interactions [7]. DNase I prefers to cut the DNA within some specific sequence context and avoids cutting within others. The study was conducted by evaluating the average DNase I cleavage in cellular samples and naked DNA around different TF binding motifs. In several cases, the cleavage patterns were highly similar despite the fact that there were no proteins present in the naked DNA sample. This clearly indicates that the intricate high resolution patterns are not informative about TF-DNA interactions in all cases and high attention should be paid when the high resolution signal can be used to discriminate real TF binding.

Figure 3.3 shows the average DNase I patterns at P53 and CTCF binding sites. For P53 the clearly distinct one nucleotide resolution pattern with two peaks at specified locations is actually produced by the sequence bias as the patterns in the naked DNA sample and the cellular sample are highly similar.

However, in some cases the high resolution pattern seems to be highly informative as it differs from the pattern in the naked DNA samples. This is the case for protein CTCF which has a highly distinct one nucleotide peak just outside the actual binding site in the cellular sample but the peak is absent in the naked DNA sample. In such case the high resolution DNase I data should provide valuable information about the TF binding and this information should be utilized in binding predictions.

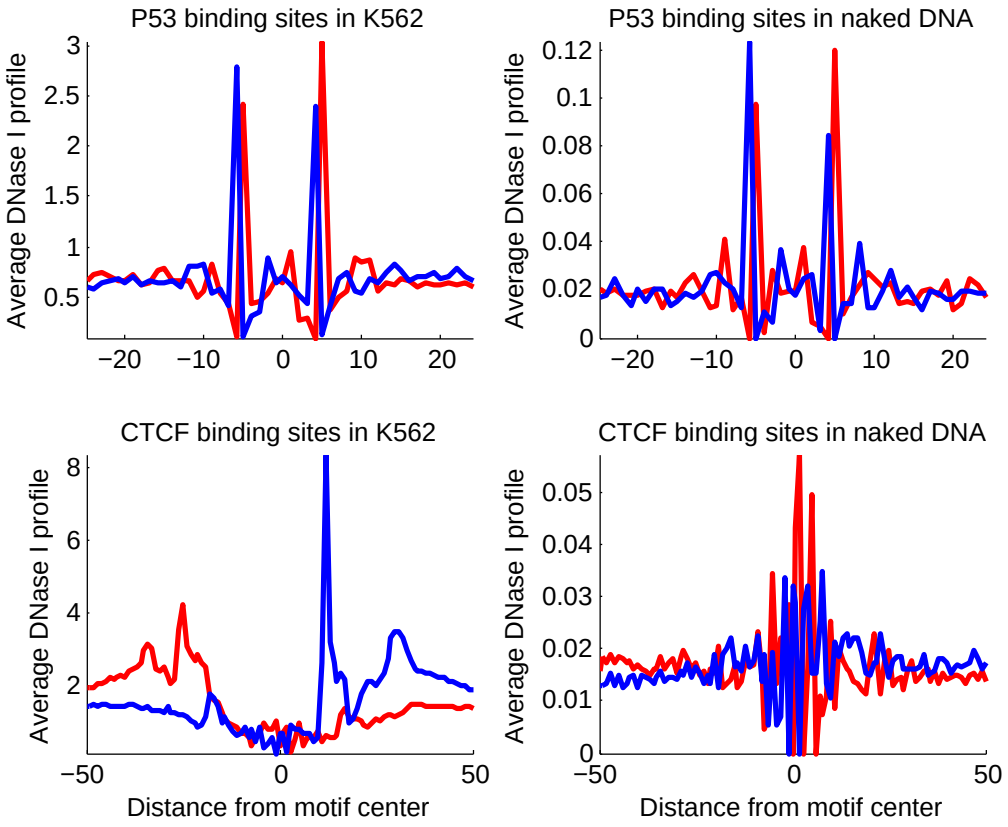


Figure 3.3: The average DNase I cleavage is shown for P53 and CTCF in cellular and naked DNA samples. Different colours indicate the DNase signal on each DNA strand (red=plus strand, blue=minus strand). The DNase I cleavage pattern at P53 binding sites in celltype K562 resembles closely to the pattern observed in naked DNA samples. The DNase signal for CTCF is clearly different from the observed sequence bias. Data for this plot is taken from [7].

Chapter 4

Materials

4.1 ChIP-seq data: ENCODE

ChIP-seq data is used in this work to determine which sites are actually bound by transcription factors. We chose to use cell type K562 data from the ENCODE project because the cell type has the highest number of TFs mapped and deeply sequenced DNase I data is also available. In this work data for 57 sequence specific transcription factors are used for celltype K562. Additionally data for 31 TFs in cell type HepG2 is used for evaluating how well the methods developed in this work generalize to new cell types (Section 6.8). The data sets used in this work are freely available through the ENCODE project page and the filenames are listed in Appendix A.

4.2 DNase I data: ENCODE

The deeply sequenced DNase I data mentioned in Section 3.1 was published for 41 cell types in 2012 [22]. The datasets from [22] (ENCODE track name UwDgf) were downloaded from the ENCODE web page for cell types K562, HepG2, SKMC and NHDF-Ad. The datasets include the raw reads, DNase I hotspots and the DNase I signal. The DNase I hotspots are regions of high DNase I activity found by the hotspot algorithm [8].

The protocol has developed afterwards [7] and the experiment has been conducted to additional cell types. In the original publication, the average sequencing depth was 273 million reads per cell type. The quality and sequencing depth of these datasets had quite large variation but the very high sequencing depth was speculated to make these differences unimportant.

4.2.1 Data preprocessing

The reads obtained from the DNase I experiment (Section 3.1) have to be subjected to computational processing steps in order to gain information about the DNase I activity. The end product of the computational steps is a genome-wide DNase I signal which reports the number of DNase I induced cuts for each nucleotide position. The processed DNase I signal is available through the ENCODE web page but this processing pipeline had to be implemented in this work because there were discrepancies in the ENCODE data sets (Section 6.1).

The standard processing pipeline as implemented in this work is presented in Figure 4.1. This pipeline is analogous to the ENCODE processing steps. First the raw reads are aligned to the reference genome (hg19) using *Bowtie (0.12.8)* [15] using parameter set (–phred64-quals -n 3 -v 3 -k 2). The aligned reads aligning to different strands are separated using *SAMtools* [17]. After separating the reads aligning to each strand we count the number of reads starting from each nucleotide position with ENCODE binary *align2rawsignal*. The result of these steps is two files containing the DNase I signal for both DNA strands.

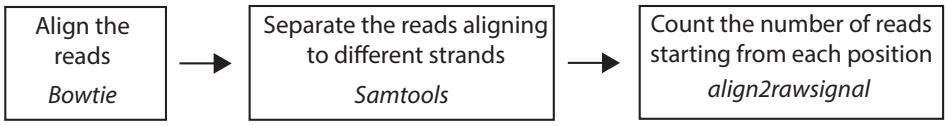


Figure 4.1: The standard DNase I data processing pipeline as implemented in this work.

The reason for separating the reads aligning to different strands is to account for the 5' nature of the DNase I experiment. If we calculate the number of reads starting from each position without taking the strandedness into account, the DNase cuts would not be correctly assigned to each nucleotide. This is shown in Figure 4.2. The DNase molecule cuts between the 3rd and 4th nucleotide in both strands resulting in reads AGC and TGG, but the DNase signal gets attributed to different nucleotides. This is corrected by shifting the reverse strand reads one nucleotide to 5' direction.

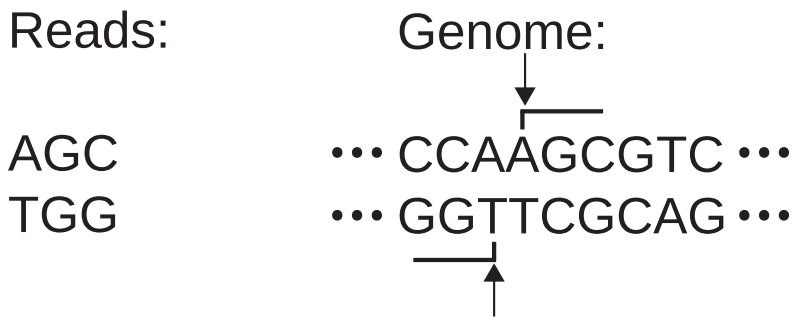


Figure 4.2: The reads AGC and TGG are the result of DNase I cleavage between the 3rd and 4th nucleotide, but the cut is attributed to different nucleotides if we do not shift one of the reads 1 bp to 5' direction

4.3 Candidate binding sites

Candidate binding sites are defined as genomic sites that contain a binding motif for a transcription factor. The candidate binding sites are found by scanning different genomic locations using a set of predefined PSFM models. The PSFMs used in this work for each TF are the primary (canonical) binding motifs reported in [30]. For the TFs that were not included in the article the models were searched from the TRANSFAC database [32]. This set of PWMs were used to scan different genomic regions using the motif scanning software FIMO [6] using p-value threshold $p = 10^{-5}$.

The motif scanning was conducted to three different sets of genomic locations: ChIP-seq regions, hotspots and random locations. The candidate binding sites were then divided into truly bound (positive) sites and unbound (negative) sites according to the list below.

- Motif within a ChIP-seq peak = Truly bound sequence (positive set)
- Motif not within a ChIP-seq peak = Unbound sequence (negative set 1)
- Motif not within a ChIP-seq peak but within a hotspot region = Unbound sequence (negative set 2)

The ultimate goal of this work is to find a way to discriminate between the positive set and the two negative sets. All of the candidate binding site data sets are further divided to training and testing sets: the training set is used for training the models and the testing set is used for evaluating the

discriminatory performance of the models. The positive testing sets contain 200 and negative testing sets 1000 candidate binding sites. The rest of the sites are used in the model training.

The DNase signal from the candidate binding sites is extracted using ENCODE binary *bigWigToWig*. The extracted signal covers the candidate binding site and 100bp up- and downstream. Each instance in the modeling spans therefore $200 + l$ basepairs, where l is the width of the binding motif.

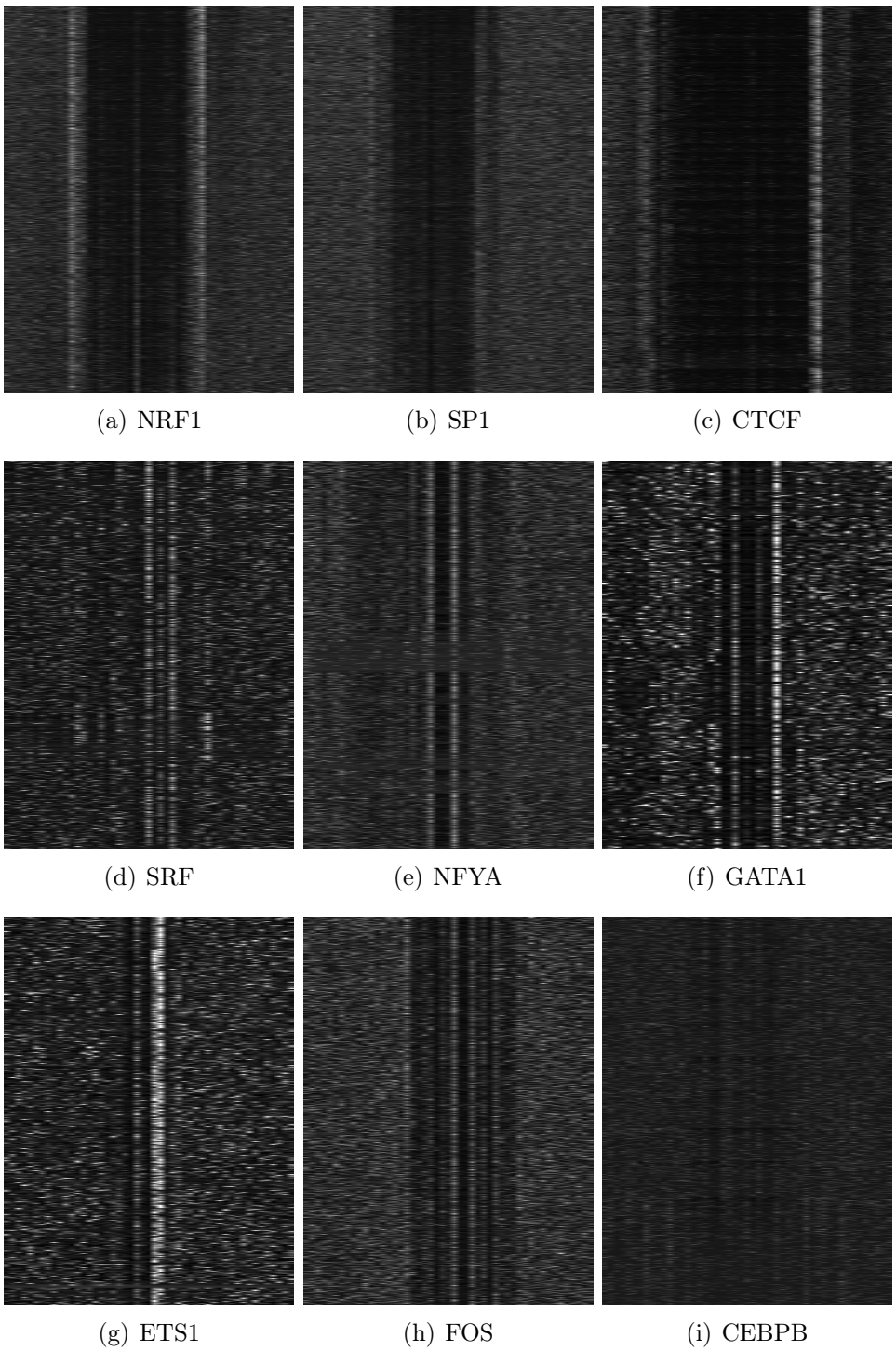
The strand from which the motif can be found has to be taken into account. The coordinates from the reverse strand motif instances have to be shifted one bp downstream for the same reason reverse stranded reads have to be shifted. The signal has to be also reversed to correctly account for the strandedness and motif orientation.

4.4 Illustrative examples from the data.

The data from candidate binding sites is shown in Figure 4.3 for nine different TFs. The rows of these heatmaps are the 50bp windows centered at candidate binding sites withing ChIP-seq peaks and the colour indicates the DNase I activity within each position (white=the highest activity, black=no activity). Numerous observations can be made from these figures. Similar observations can be made for other TFs included in this study. The heatmaps for remaining 48 factors is presented in Appendix B.

For many TFs, the data follows the canonical form of the TF footprint: the DNase activity right in the middle is low and the flanking regions exhibit higher activity. This is particularly clear for the upper row transcription factors NRF1, SP1 and CTCF. This form of TF footprint reflects the idea that TF binding to the DNA protects the underlying DNA from DNase I cleavage as explained in Section 3.1. In addition to the canonical footprint pattern NRF1 has single nucleotide wide peaks flanking the candidate binding site and CTCF has a single nucleotide peak downstream from the binding site.

For some TFs such as NFYA, GATA1 and FOS the canonical form of the TF footprint is mixed with strong one nucleotide peaks within the candidate binding sites. This might be the result of exposure to DNase I cleavage or the intrinsing sequence bias of the DNase I molecule. For SRF and ETS1 only nucleotide resolution peaks can be found. There is no clear decrease in DNase activity directly within the candidate binding site. For CEBPB, the DNase I pattern seems to tell nothing. From the data we can nevertheless see that the DNase I activity is low throughout the binding sites. Similar low DNase I activity candidate binding sites within ChIP-seq peaks can be found for proteins MAFF, MAFK and ZNF274 (Appendix B).



(a) NRF1 (b) SP1 (c) CTCF

(d) SRF (e) NFYA (f) GATA1

(g) ETS1 (h) FOS (i) CEBPB

Figure 4.3: Heatmaps displaying the DNase I activity within the candidate binding sites for nine TFs.

Chapter 5

Methods

5.1 PSFM modeling

Section 2.2 explained that PSFM binding modeling is the state-of-the-art method for computationally determining TF binding sites. This section explains the mathematical details for calculating a binding score for a nucleotide sequence. The binding score reflects the affinity of the TF to a specific sequence.

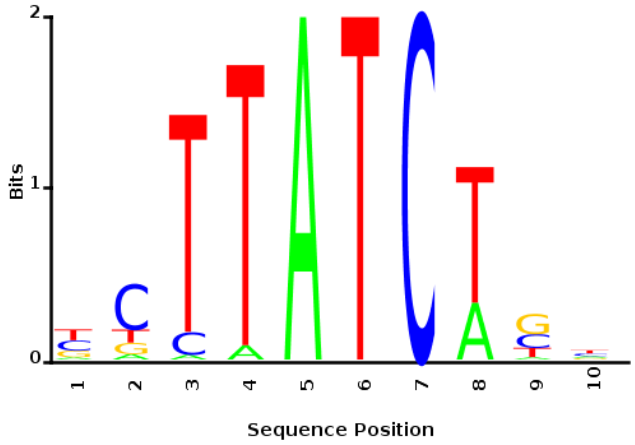
When the PSFM models are constructed a set of known bound sequences is used to determine the probability of each nucleotide in each position. The sequences are laid on top of each other (aligned) and the number of occurrences of each base is calculated in each position. Typically, a pseudocount is added to the counts in this procedure to avoid zeros. Pseudocounts can be assigned in many ways: the simplest way is to add one to each of the counts. The counts are then turned into probabilities. Figure 5.1 shows the PSFM model and the sequence logo visualization for protein GATA1. This PSFM does not include pseudocounts as can be seen from the zero entries in the matrix.

The likelihood of a sequence produced from the PSFM model can be seen in Equation 5.1:

$$\mathcal{L}_{PSFM} = \prod_{i=1}^{l} p_i(B_i) \tag{5.1}$$

where i is the position in the nucleotide sequence, l is the length of the matrix model, B_i is the observed base and $p_i(B_i)$ is the probability of the observed base in position i given by the PSFM.

In order to tell if the sequence appears only by chance, we have to compare this likelihood with the probability of the sequence produced by the



$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{pmatrix} 0.08 & 0.07 & 0.02 & 0.05 & 1 & 0 & 0 & 0.30 & 0.05 & 0.19 \\ 0.33 & 0.60 & 0.09 & 0 & 0 & 0 & 1 & 0 & 0.30 & 0.29 \\ 0.22 & 0.15 & 0 & 0 & 0 & 0 & 0 & 0 & 0.44 & 0.18 \\ 0.37 & 0.18 & 0.89 & 0.95 & 0 & 1 & 0 & 0.70 & 0.21 & 0.34 \end{pmatrix} \end{array}$$

Figure 5.1: The PSFM for GATA1 and the corresponding sequence logo visualization

background distribution. The most widely used background distribution is a zero order distribution in which the probabilities of the nucleotides are the same in each position. The likelihood can be calculated with Equation 5.2.

$$\mathcal{L}_{Background} = \prod_{B \in (A,T,G,C)} p(B)^{n(B)} \tag{5.2}$$

where $p(B)$ is the probability of base B in the background distribution and $n(B)$ is the number of bases B in the sequence.

The likelihoods are used to calculate the negative log-likelihood ratio using Equation 5.3:

$$\mathcal{S}_{\mathcal{PSFM}} = -ln\left(\frac{\mathcal{L}_{Background}}{\mathcal{L}_{PSFM}}\right) \tag{5.3}$$

where ln is the natural logarithm. This ratio can be interpreted as a binding score ($\mathcal{S}_{\mathcal{PSFM}}$) which tells the affinity of the TF to the given sequence. The higher this value is, the more likely the TF will bind the sequence. These scores can be turned to p-values using dynamic programming [27]. The p-values are often used in motif scanning software such as FIMO (Section 4.3).

5.2 Multinomial distribution

Multinomial distribution can be used to model the probabilities of combinations of discrete counts. The counts (summing to n) are the results of n independent tests each of which will lead to the success of one categorical variable with a fixed probability. The probabilities for all the categories sum to one. The probability for each count combination can be calculated with the probability mass function (Equation 5.4) where the xs are the counts in each class and the ps are the corresponding probabilities.

$$\frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \tag{5.4}$$

The multinomial distribution can be applied to DNase I hypersensitivity data by treating the nucleotides within the candidate binding region as categorical classes. The DNase counts in these coordinates are the variables. In this work, a window of 50bp around the candidate binding site is used. This way the actual binding site is covered and the flanking nucleotides are also included in the modeling.

The multinomial distribution is used to calculate the likelihood of the candidate binding site and the likelihood is considered as the binding score. Because the number of counts in each window is not equal, the binding scores are not immediately comparable. For this reason a simple sampling scheme was devised in which the counts in each window sum up to the same fixed number:

- Estimate the underlying multinomial distribution by normalizing the counts in the window to sum up to one.
- Make a sample of 100 trials out of the estimated distribution.

The likelihoods of these samples are comparable because the likelihood differences are no longer caused by the difference in the total number of counts.

5.3 Logistic regression

Logistic regression can be used to map real values to probabilities using the logistic function (Equation 5.5).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5.5}$$

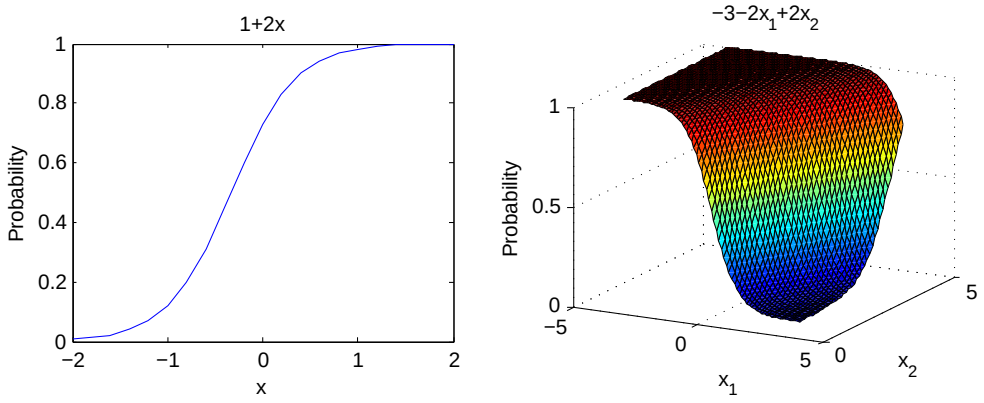


Figure 5.2: The logistic function using one or two variables.

The logistic function produces always values between one and zero, so it can be easily interpreted as a way to represent probabilities. This is illustrated in Figure 5.2, which shows one and two-dimensional logistic functions.

If the input variable is a linear function, logistic regression is a form of a generalized linear model. In this case the function values or probabilities can be represented as in Equation 5.6:

$$f(x) = \frac{1}{1 + e^{-ax^T}} \tag{5.6}$$

where a is the coefficient vector and x is the vector corresponding to the variables. This is illustrated in Figure 5.2. In the left figure the coefficient matrix is [1 2] and in the right figure it is [-3 -2 2] as can be seen from the corresponding titles. With slight modification to Equation 5.6 we get Equation 5.7 :

$$log\left(\frac{f(x)}{1 - f(x)}\right) = ax^T = \beta_0 + \beta_1 x_1 + \beta_1 x_2 \ldots \tag{5.7}$$

where a is the coefficient vector $[\beta_0,\ \beta_1,\ \beta_2 \ldots]$ and x is the vector corresponding to the variables $[1,\ x_1,\ x_2 \ldots]$. Note that the first element of x contains a 1 which is the multiplier for the constant term β_0.

Logistic regression can be applied to DNase I hypersensitivity data by treating the DNase I induced counts as variables, as explained in Section 5.2. Logistic regression has a clear advantage over the multinomial modeling as any real number variables can be straightforwardly added to the modeling. In this work, the scores from PSFM modeling are used as additional information.

5.4 Area under the ROC curve

The receiver operating characteristic (ROC) curve is a way to describe the performance of a classifier in a binary classification task. In the classification task, each data point is given a score and the data is classified into two classes using a threshold value. The true classes of the data points have to be known in order to calculate the ROC curve. The threshold of the classifier is varied and using the true classes and the estimated classes we can calculate true positive rate (TPR) and false positive rates (FPR) which are given in Equation 5.8, where TP=True Positives, P=Positives, FP=False Positives, N=Negatives.

$$TPR = \frac{TP}{P} \qquad FPR = \frac{FP}{N} \tag{5.8}$$

The curve can be plotted as a two-dimensional plot where the true positive rate is on the y-axis and the false positive rate is on the x-axis. The area under this curve (AUC) gives the probability of a positive instance scoring higher than a negative one, and can therefore be considered as a performance measure in the classification task. Higher AUC-value indicates a better classifier.

In this work, AUC is used as a model performance statistic. AUC metric gives the probability for a truely bound site to score higher in the modeling than an unbound site.

5.5 Feature selection using cross validation

Feature selection is a form of dimensionality reduction in which a subset of variables is chosen for modeling purposes. In this work this means choosing the relevant nucleotide positions. Typically the number of feature subsets is too large for performing an exhaustive search for the best subset. For this reason heuristic methods have to be used. In this work a greedy backward selection method using cross validation is implemented for selecting the features for single nucleotide resolution multinomial and logistic regression models.

The backward search utilizes cross validation to evaluate the performance change caused by each potential feature selection operation. In cross validation the model training data is further divided in partitions for model training and evaluation. In cross validation, the data division can be conducted many ways. In this work a fixed number of instances is always sampled for model evaluation and the rest are used as training data. The data is divided

Algorithm 1 Backward selection

```
1: Start with N features
2: for each remaining feature do
3:     Remove the feature temporarily.
4:     for 30 times do
5:         Divide the data in two partitions.
6:         Train the model using the first partition.
7:         Evaluate the model using the second partition.
8: Remove the worst feature
9: goto line 2
```

and the effect of removing each feature is evaluated 30 times and the average of the performance is taken for reducing the randomness in this search caused by the division of the data in training and evaluating partitions. In each iteration the feature whose removal results to the highest increase or smallest decrease in the performance evaluation is removed. In the following pseudocode presentation this feature is dubbed *the worst feature*.

5.6 Binning as feature extraction

In the context of this work data binning means combining the data into wider windows instead of using the highest possible resolution. Instead of looking at the DNase counts for each nucleotide, the sum of counts in adjacent positions is treated as a variable. Binning can be beneficial if the data is noisy or if there is lack of coverage.

Data binning is a feature extraction operation in which the input data from multiple different positions is transformed into a integer count. The binning can be assigned in combinatorially many ways. The bins can be of different sizes and they can cover different regions with respect to the candidate binding site.

5.6.1 MILLIPEDE

MILLIPEDE is a method for predicting DNA-TF interactions using DNase hypersensitivity data [18]. The method is used for classifying candidate binding sites to either bound or unbound using a logistic regression model. The variables in the logistic regression are the log-transformed counts of DNase induced cuts in each bin around and within the candidate binding site.

The method uses six different ways for assigning the data into the bins.

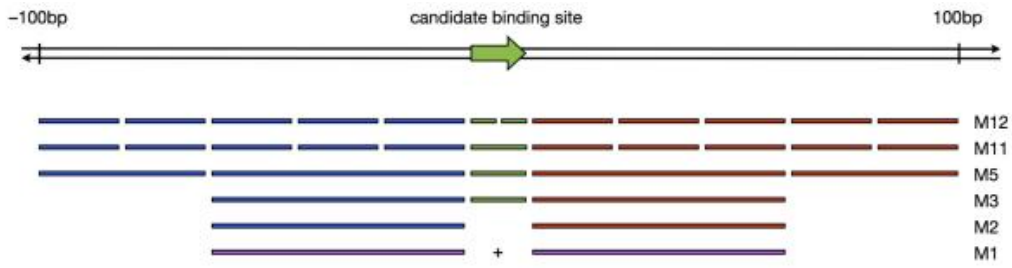


Figure 5.3: MILLIPEDE assigns the data in bins in six different ways as seen in the figure.

Figure 5.3 shows how the reads are aggregated in the different MILLIPEDE models. The data directly within the candidate binding site is modelled as a separate or two separate bins, and the flanking data is modeled using wider bins of different sizes.

The probability for binding for each candidate binding site can be solved from Equation 5.9:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_i \beta_i D_i + \beta_{PSFM}\mathcal{S}_{\mathcal{PSFM}} \tag{5.9}$$

where p is the probability of binding, D_i is the log-transformed number of reads in the bin i, S_{pwm} is the score in the PWM-modeling and β are the coefficients in the model. This formula is identical to Equation 5.7.

MILLIPEDE is used as a comparison to the models developed in this work. The method has been shown to outperform the CENTIPEDE method which is widely used for predicting TF-DNA interactions using DNase I data [18, 25].

5.6.2 BinDNase

In this work, a greedy backward search type machine learning method is implemented to find (and extract) optimal features from the DNase I data for each TF. The greedy backward search is identical to the feature selection process described in Section 5.5 but in this case the operation conducted after each iteration is different. At the beginning of the search the nucleotides in the candidate binding site and the flanking 10bp regions are treated in one nucleotide resolution, and the more distal nucleotides are initially in ten nucleotide wide bins. These initial bins are then merged in the search to achieve more predictive power. In each iteration step two bins are merged.

Algorithm 2 BinDNase bin selection

1: Initialize the bins.
2: **for** each remaining pair of adjacent bins **do**
3: Merge the bins temporarily.
4: **for** 30 times **do**
5: Divide the data in two partitions.
6: Train the model using the first partition.
7: Evaluate the model using the second partition.
8: Merge the two bins resulting in best performance.
9: **goto** line 2

This is presented as a pseudocode below and a schematic presentation with two iterations is presented in Figure 5.4.

In each iteration the algorithm performs the particular feature extraction operation that leads to the best prediction performance. The prediction performance is evaluated using cross validation by dividing the training data into two sets (both positive and negative sets). The model is trained using one of the sets and the model performance is evaluated using the other set by making predictions and calculating the area under the curve (AUC) metric. This cross validation step is conducted 30 times for each feature extraction operation to reduce the variation caused by random division of the data into training and test sets in the model training.

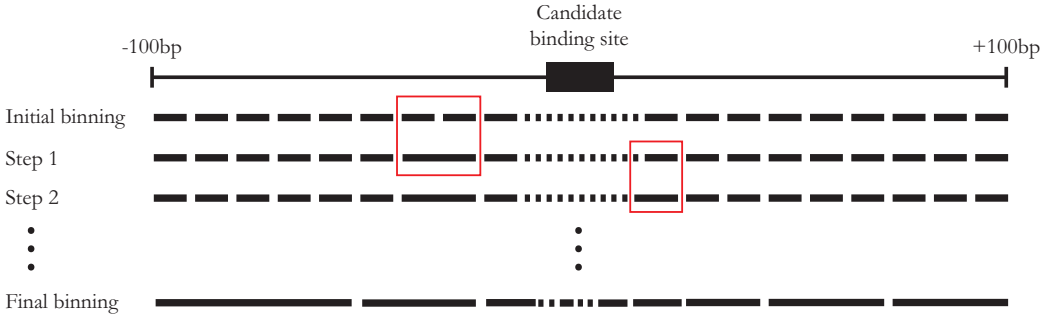


Figure 5.4: A schematic presentation of how the optimal binning is found. In the initial stage the candidate binding site and 10bp flanking regions are modeled using 1bp resolution. The flanking regions 11-100bp up and downstream from the candidate binding motif are modeled using 10bp bins. In each step two bins are merged. In this figure the first step of the algorithm merges two bins upstream of the binding site. In the second step one 1bp wide window is merged to the right flanking bin. The final binning in this illustration includes wide bins at the flanking regions and narrower bins at the binding site.

Chapter 6

Results

6.1 Standardized DNase-seq data preprocessing is prerequisite for single nucleotide level analysis

In all modelling the data should be preprocessed in a standardized way. Previous articles have proposed slightly different processing steps for DNase-seq data (see e.g. [22, 24]). In the DNase I data special attention should be paid on the following processing steps, as explained in Sections 4.2.1 and 4.3.

- Reverse strand reads should be shifted 1 bp to 5' direction (or alternatively forward strand reads should be shifted 1 bp to 5' direction). This shifting acknowledges the fact that DNase I cleaves the DNA between two consecutive nucleotides. With a single base pair shift on either of the strands, the DNA cut sites are contributed consistently to a single nucleotide.

- Orientation of TF binding motifs should be taken into account.

Discrepancies in these steps lead to differences in the data and therefore affect any downstream analysis. For example, some of the digital DNase I data sets which are available on the ENCODE project page have preprocessing differences between cell types. Figure 6.1 shows the average DNase I cleavage profiles for protein JUN in four ENCODE cell types: NHDF-Ad, SKMC, K562, and HepG2. We noticed that reverse strand reads are not consistently shifted between these four cell types (compare Figures 6.1 a)-b) with c)-d)). While the overall pattern of wider DNase hotspot remains practically unaffected in cell types K562 and HepG2, the nucleotide resolution cleavage patterns in the JUN interaction site is significantly distorted due

to non-standardized data preprocessing. Consequently, if these discrepancies are not properly corrected, TF binding prediction methods which use high resolution DNase-seq data, such as BinDNase, fail to generalize between cell types. Starting from the raw reads and reprocessing the data carefully using a unified preprocessing pipeline (Section 4.2.1) makes differences between cell types disappear (Figure 6.1 e)-f)). We observe a similar effect for the majority of TFs. For TFs for which no distinct single-nucleotide resolution signal can be observed, the distortion caused by differences in preprocessing steps is not obvious. We conclude that careful and highly standardized data preprocessing is essential for detailed analysis of DNase-seq data.

6.2 Transcription factors can be clustered using DNase I data

The binding sites of TFs belonging to similar structural groups exhibit similar DNase I profiles. This can be seen from the clustering presented in Figure 6.2, where many proteins belonging to the same structural groups cluster together in a clustering using the average DNase I signal within candidate binding sites. Additionally, DNase profiles for six TFs belonging to the same structural group (HLH) is presented in Figure 6.3. The figure illustrates the degree of similarity observed in the DNase signal for these TFs.

The clustering is done by comparing the average DNase signal around and within the candidate binding sites of each TF. A distance is calculated between each pair of transcription factors by evaluating the correlation of the DNase signal in a 30 bp sliding window centered at the candidate binding site. First the average signal is calculated using the data from candidate binding sites and a 50 bp window is taken around the site. When comparing two TFs, a 30 bp window from the middle of the average profile of one of the TFs is compared with every possible 30 bp window in the 50 bp window of the other TF. For each possible position, the correlation is calculated. Only the maximum correlation is taken into account, because we consider that the TF can bind only to one location within each candidate binding site. As correlation represents similarity between the two signals, $1 - Cor$ can be interpreted as a distance measure. This distance measure is used to make the clustering using standard hierarchical clustering which uses the shortest distance between the clusters as the distance (*linkage*-function with default settings in MATLAB).

The clearest cluster in Figure 6.2 is the cluster formed by JUNB, JUN, JUND, FOS and FOSL1. These proteins form the dimeric activation pro-

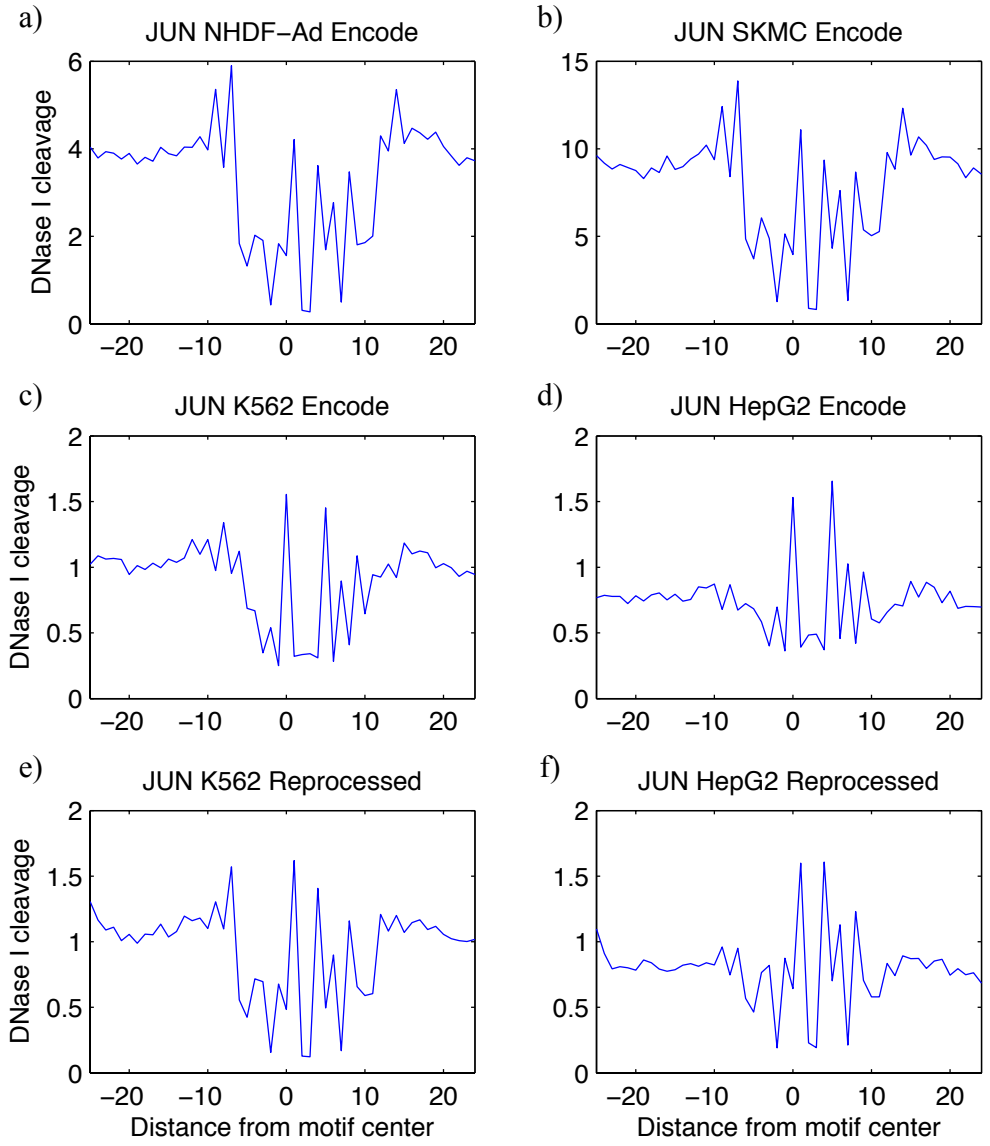


Figure 6.1: Average DNase I cleavage profiles for protein JUN in four ENCODE cell types: a) NHDF-Ad, b) SKMC, c) K562, d) HepG2, e) K562 reprocessed, and f) HepG2 reprocessed. For each cell type, the average DNase-seq signal at nucleotide resolution centered at JUN motif overlapping DNase-seq hotspot is shown. The celltypes K562 and HepG2 exhibit clearly distinct average pattern in c) and d). Careful preprocessing of the data makes these unexpected cell type specific differences disappear as shown in e) and f) and is essential for nucleotide resolution analysis of the DNase I data.

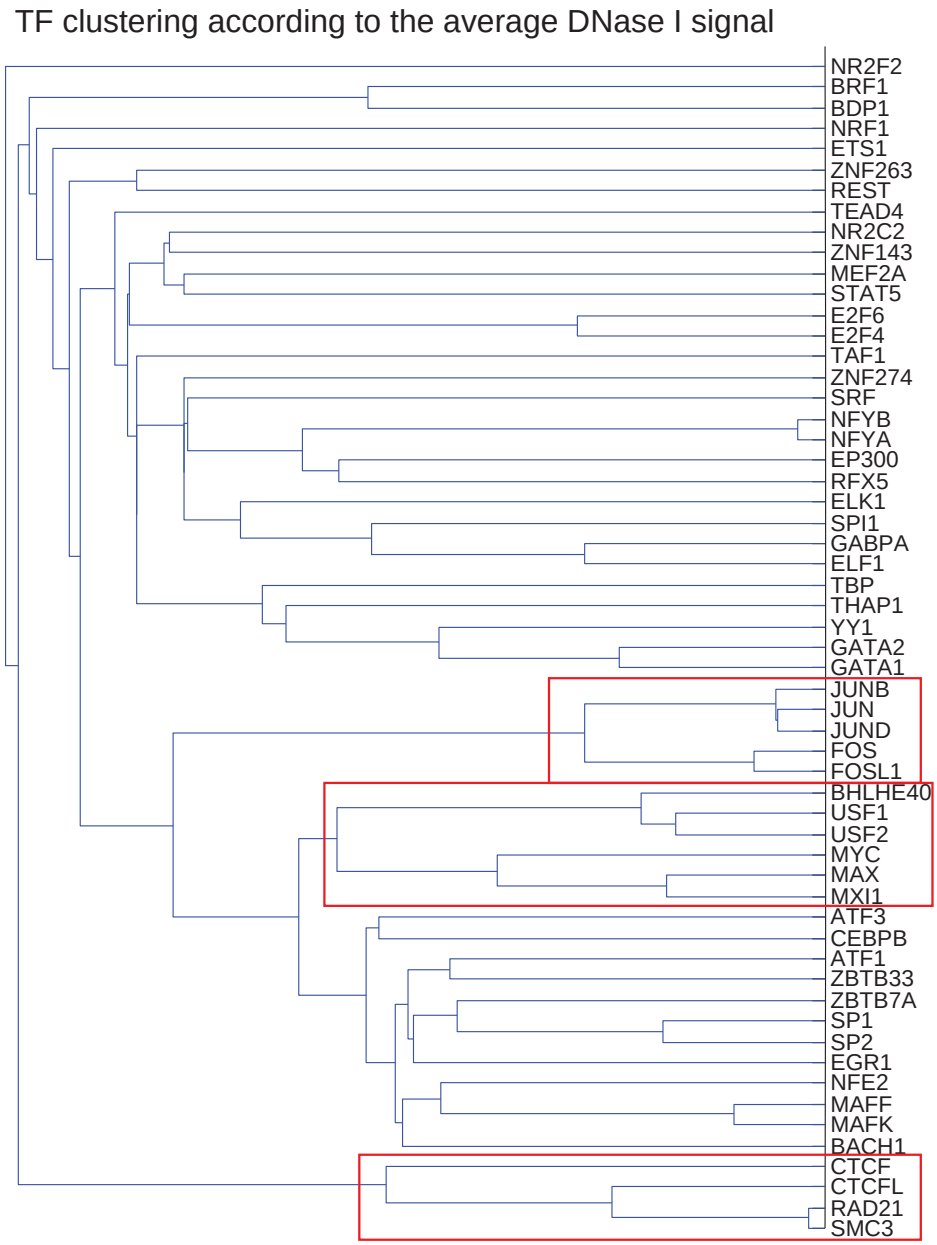


Figure 6.2: The clustering of transcription factors using the average DNase I signal in the motif instances within ChIP-seq peaks.

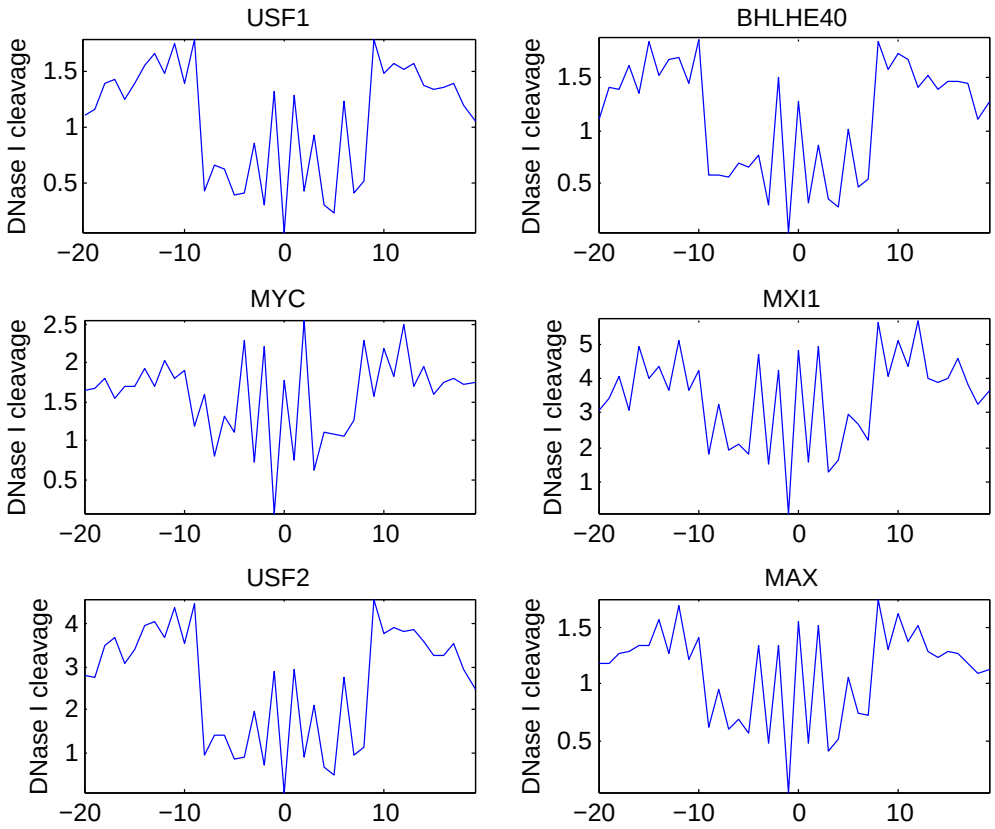


Figure 6.3: The average DNase I signal for six transcription factors belonging to the basic HLH structural classes. The canonical footprint shape can be seen the best for TFs USF1, USF2 and BHLHE40. The depletion of DNase I cleavage is not so clear in the middle of TF-DNA interaction for TFs MYC and MXI1. The average signals for all these TFs are highly correlated.

tein 1 (AP-1). AP-1 is constructed using proteins from JUN, FOS and ATF protein families [12]. Although the different AP-1 transcription factors differ slightly, they recognize similar sequences.. Therefore the binding motif used for JUN and FOS proteins used in this work is the same (AP-1 motif) which might contribute to the TFs clustering together, because the sequence contributes to the shape of the DNase signal via the intrinsic sequence bias.

Other notable TFs grouping together include CTCF, CTCFL, RAD21 and SMC3 which are all somehow involved in chromatin modulation. The canonical binding motif for these TFs is the same (CTCF-motif), which might contribute to the fact that these TFs cluster together, similarly as in the AP-1 case.

Another interesting group is formed by BHLHE40, USF1, USF2, MYC,

MAX and MXI1 which are all members of basic helix-loop-helix leucine zipper and related helix-loop-helix structural families. These both structural families are members of the biggest transcription factor superclass which includes basic structural binding domains (Section 2.1). The binding motifs for these TFs are similar but not identical. The average cleavage pattern for these TFs can be seen in Figure 6.3. Note that the single-nucleotide patterns are in slightly different positions. The position with almost complete depletion from cleavage right in the middle of the motif (0 or -1 bp from the center) differs between the factors, but this is taken into account in the distance measure by using a sliding window and the TFs will cluster together.

These results show that there are clear similarities in the DNase signal of related TFs. This might prove to be problematic in future studies, because if similar TFs bind the same DNA sequence and DNase profiles for the binding sites of these TFs resemble each other, the TFs cannot be distinguished by using only DNase I hypersensitivity data.

6.3 Logistic regression outperforms the multinomial method

The use of high-resolution DNase I data using single nucleotide resolution modeling was studied using the multinomial method and logistic regression as described in Sections 5.2 and 5.3. Feature selection was used for finding the optimal nucleotide positions for the modeling. The prediction performance comparison for these two approaches can be seen in Figure 6.4 a). It can be seen that logistic regression works better than the multinomial method as most of the triangles representing the TFs are below the diagonal.

The difference in the prediction performance can be explained by the fact that PSFM-modeling scores can readily be used as additional information in the logistic regression model. This is also demonstrated in Figure 6.4 b), which compares the prediction performance of the logistic regression method with and without using the PSFM information. For most TFs the prediction performance is almost identical which means that for those TFs the PSFM-scores do not help in discriminating the truely bound sites. However there are many cases in which the PSFM-score is informative and can help improving the prediction accuracy.

The multinomial method also poses some difficulties as the likelihoods are not immediately comparable and the sampling scheme described in Section 5.2 has to be used. This sampling scheme is not necessarily optimal way to utilize the properties of the multinomial distribution as it definitely increases

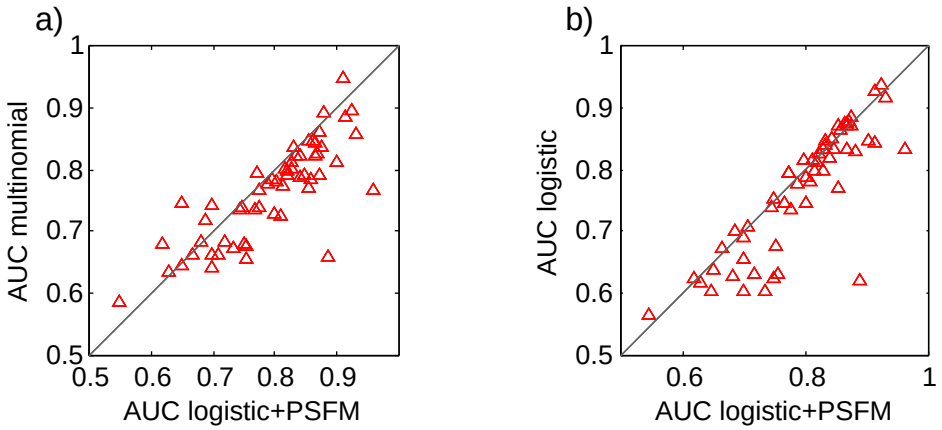


Figure 6.4: The multinomial model is compared against the logistic regression model in a). Figure b) shows the prediction performance of the logistic model with and without using the PSFM-scores as additional information.

noise due to the randomness in the sampling process. For these reasons, the logistic regression seems to be more suitable modeling framework for TF-DNA predictions using DNase-seq data than the simple multinomial method.

6.4 DNA binding should be modelled separately for each TF

After training the different models as described in Chapter 5 using the data as explained in Section 4.3, we observed that the selected features as well as the actual prediction models differ greatly between TFs. Some of the BinDNase models are visualised in Figures 6.5 and 6.6. In the figures the upper panels show the average DNase I cleavage at the candidate binding sites and the coloured bars indicate the optimised feature selection (the bins found by BinDNase) and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient.

As can be seen in Figure 6.5, for some TFs, such as ATF1 and SP1, the canonical definition of DNase I footprint is adequate as the feature selection algorithm finds a model in which the reads falling in the central bins decrease the binding score and the reads falling in the flanking nucleotides increase the binding score (Figures 6.5 a) and b)). This could be also speculated even before the model building, because the average cleavage profile does not show any clear high resolution patterns.

We also observed that not all strong single nucleotide cleavage patterns

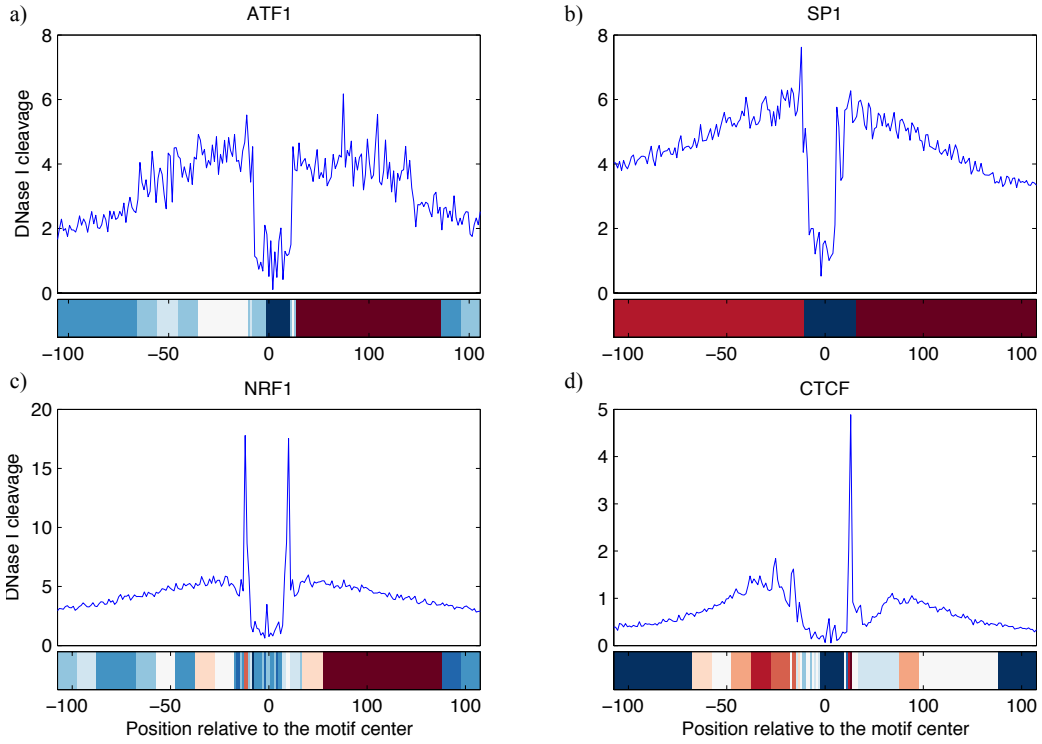


Figure 6.5: The upper panels show the average DNase I cleavage centered at the TF binding motifs. The coloured bars indicate the optimised feature selection and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient. a) ATF1, b) SP1, c) NRF1, and d) CTCF.

present in data are important for discriminating true TF binding sites from random unbound motif locations. For example, NRF1 protein has three evident single nucleotide resolution cleavage sites but only the one to the left from the motif is associated with a high positive regression weight by the feature selection algorithm (see Figure 6.5 c)). Note that reads falling to the flanking region on the right are weighted more than the single-nucleotide resolution pattern. Thus, some of the intricate patterns in DNase-seq data seem unimportant for discriminating between real TF binding and noise and efficient feature selection is needed to identify the relevant DNase I cleavage patterns. On the other hand, some of the single nucleotide resolution patterns are highly informative. For example, the DNase I signal for protein CTCF (Figure 6.5 d)) contains a highly stereotypical peak on the right hand side of the binding site. In the optimised feature selection this nucleotide is treated as a single nucleotide and the coefficient in the logistic regression model is high. Moreover, the cleavage pattern in this exact position relative to CTCF

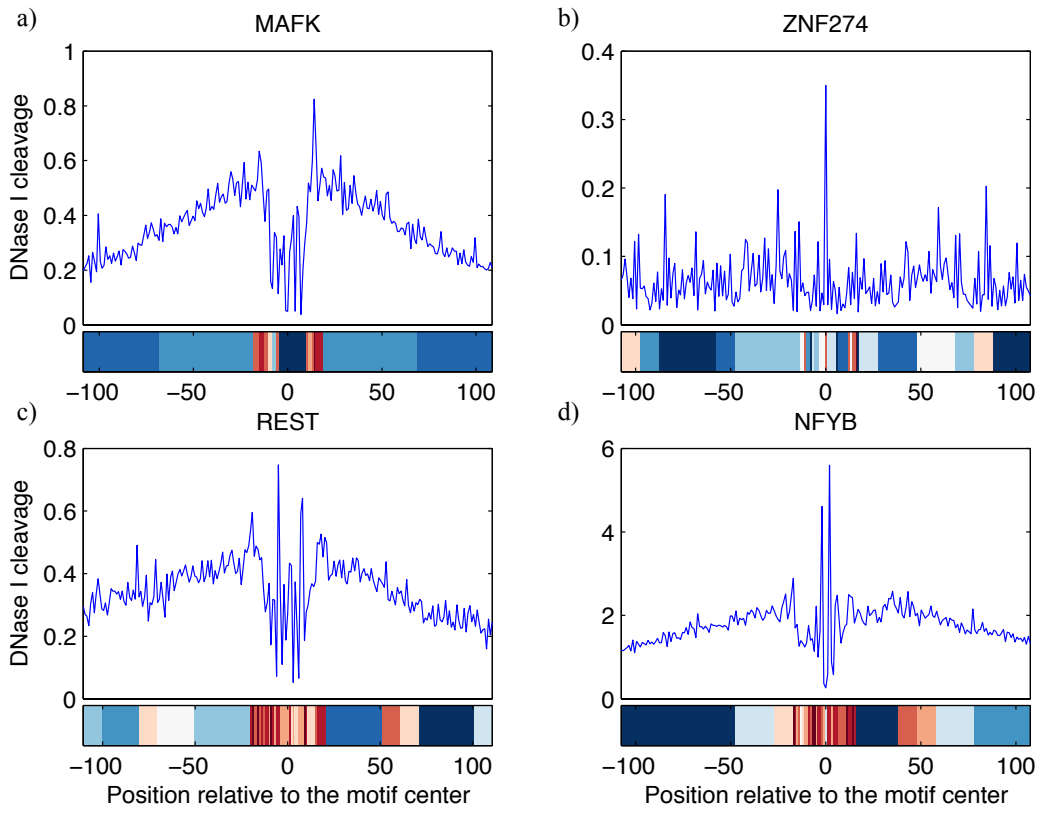


Figure 6.6: The upper panels show the average DNase I cleavage centered at the TF binding motifs. The coloured bars indicate the optimised feature selection and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient. a) MAFK, b) ZNF274, c) REST, and d) NFYB.

motif has previously been reported to differ from the intrinsic sequence bias of the DNase I molecule [7] as discussed in Section 3.2.

The models for TFs in Figure 6.6 also show interesting phenomena. For TFs MAFK and ZNF274 the binding models seem to suggest that low DNase activity within the hotspots is associated to the binding of these factors as the coefficients in the model are mostly negative (blue in the bars). The MAFK binding model seems to exhibit a traditional footprint shape as the reads falling to the narrow regions flanking the binding motif increase the binding score (red coefficient). Reads falling to farther from the binding site lower the binding prediction score systematically. For ZNF274, there is no clear pattern in the binding model. The model however includes almost exclusively negative coefficients, so it can be said that ZNF274 binding is more likely to be bound in candidate binding sites with low DNase activity.

Instead of the canonical footprint that favours a high signal in the flank-

ing regions, the models for NFYB and REST seem to emphasize the reads directly in the TF-DNA interaction site (Figure 6.6 c) and d)). This might indicate that these nucleotides are not protected from DNase I cleavage and this should be taken into account in the modeling. The pattern in the coefficients for REST is close to a reversed footprint as reads falling to flanking regions decrease the binding score.

Similar findings to those presented in this section can be found for many of the 57 TFs investigated in this study. Taken together, these results emphasize that instead of using a single TF footprint definition, the prediction models for each TF should be constructed separately.

6.5 High resolution DNase-seq analysis improves TF binding predictions

DNase I hypersensitivity at a lower resolution (i.e., for larger genomic regions) has previously been used to detect TF binding sites (see e.g. [7]). The binding score of each candidate site is then simply the aggregate counts of DNase I cuts in that larger window. We evaluated the predictive power of such a lower resolution DNase I activity method using a 50 bp window around the candidate binding site and compared that with BinDNase.

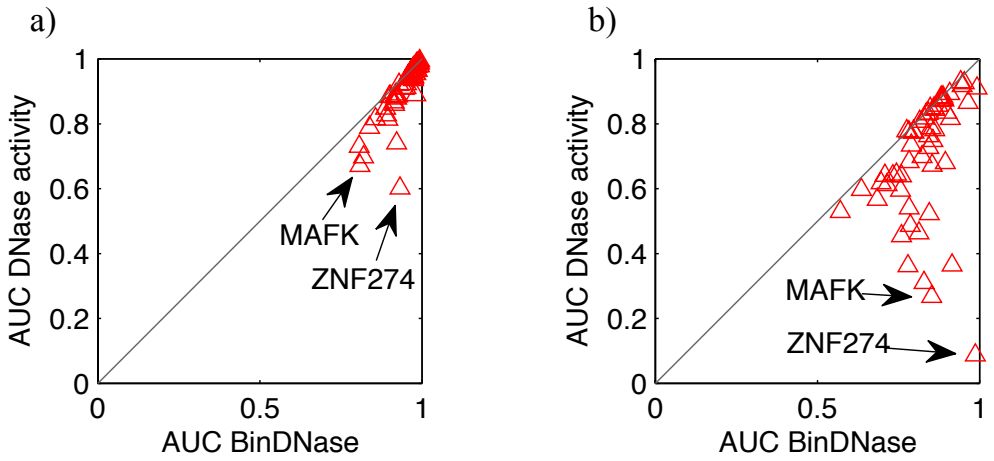


Figure 6.7: High resolution DNase-seq analysis improves TF binding predictions when compared to traditional methods. The DNase I activity predictor using a 50 bp window is compared with BinDNase using a) negative set 1 and b) negative set 2. The AUC score is computed for all 57 TFs using both methods. Selected TFs are highlighted in the figure.

In Figure 6.7 a), the TF binding prediction performance is evaluated using candidate sites from random genomic locations (negative set 1) and in Figure 6.7 b) using only candidate sites from the hotspot regions (negative set 2). For the less challenging task of discriminating real binding sites from non-binding sites which are not in hotspot regions, we observe that for many TFs it is sufficient to just quantify general DNase I activity near the binding site without using any sophisticated modeling (Figure 6.7 a)).

Although making predictions by only measuring DNase I activity is adequate, the predictions made with BinDNase are in all cases more accurate as can be seen in Figure 6.7 a). Additionally, there are TFs for which BinDNase improves the binding prediction accuracy greatly already in this scenario, including e.g. MAFK and ZNF274. The prediction models for MAFK and ZNF274 are shown in Figures 6.6 a) and b), which suggest that the relatively poor performance of the simple DNase I activity predictor can be explained by low average DNase I signal at these motif sites. BinDNase, in turn, can identify discriminatory features from DNase-seq data and increase the AUC score for MAFK and ZNF274.

Typically TF binding sites are primarily searched for in DNase I hotspot regions and, thus, performance evaluation using the negative set 2 is in practice more relevant. Results in Figure 6.7 b) show that while the DNase I activity predictor still works for some TFs surprisingly well it also fails completely for some TFs. The most notable examples include again MAFK and ZNF274 proteins for which the prediction accuracy is well below the random coin flipping. The worse than random performance can be explained by below-average DNase I activity in the MAFK and ZNF274 binding sites (Figures 6.6 a) and b)). As with negative set 1, BinDNase can identify discriminatory features from DNase-seq data and improves AUC scores for all of the 57 TFs, including MAFK and ZNF274. Interestingly, the prediction models for both MAFK and ZNF274 include high resolution features, such as a strongly positively weighted single nucleotide feature in the middle of the ZNF274 binding site (Figure 6.6 b)).

These results emphasize that DNase I hypersensitivity data is a powerful tool for identifying transcription factor binding sites genome wide. DNase I activity over a wider (50bp) region already predicts DNA binding efficiently as the DNase activity within random genomic locations is very low or non-existent. When zooming into the DNase I hotspots, the DNase I activity itself is no longer as informative as the background DNase I activity is much higher. For this reason more sophisticated modeling approaches, such as BinDNase, should be used when making predictions within the hotspot regions.

6.6 Data binning as feature extraction method improves TF binding predictions

To test if feature selection implemented in BinDNase improves TF binding predictions we also implemented methods which use DNase-seq data only at single nucleotide resolution as described in Chapter 5. We devised the methods to start with 50 nucleotides around the motif center and allow the feature selection to select the most informative features by ignoring least informative feature during each step of the feature selection process (but do not allow to combine neighbouring nucleotides into larger bins). Results in Figure 6.8 a) show that BinDNase's feature selection improves prediction accuracy results for the majority of TFs. The most notable improvement of AUC score is achieved for REST protein. In the optimal BinDNase model for REST the nucleotides in the middle are modeled at (or close to) nucleotide resolution whereas the bins at flanking sequences are much wider. Without feature selection, some of the nucleotides at the flanking sequences obtain an unnecessarily large weight and hence decrease the prediction accuracy of the nucleotide resolution model. These results suggest that for some nucleotide locations the DNase I signal does not provide single nucleotide resolution information about TF-DNA interaction and those regions should be modelled using larger bins chosen based on a feature selection method. This is expected because the DNase-seq signal at individual candidate binding sites is often noisy and part of the signal originates from the inherent DNase I sequence bias [7]. Nevertheless, we demonstrate that feature selection can identify discriminatory information from DNase-seq data.

We next compared BinDNase with the state-of-the-art discriminative method Millipede [18] (Figure 6.8 b)). Millipede has been shown to outperform the widely used CENTIPEDE algorithm [25] for predicting TF binding [18]. BinDNase performs nearly identically with Millipede for those proteins which are already predicted well by Millipede. One such protein is SP1 for which BinDNase finds a prediction model which is close to the canonical footprint model also used in Millipede (Figure 6.5 b)). However, BinDNase achieves a notable improvement for many of those proteins which Millipede does not predict well, including e.g. NFYB. Although BinDNase models the flanking regions around the NFYB motif using large size bins similarly with MILLIPEDE, the region close to the NFYB motif center is modeled using very high resolution features with high logistic regression coefficients (Figure 6.6 d)). BinDNase improves also e.g. CTCF insulator protein whose prediction model is shown in Figure 6.5 d). Previous studies have already indicated a high resolution DNase I cleavage signal on 3' end of the CTCF motif (Sec-

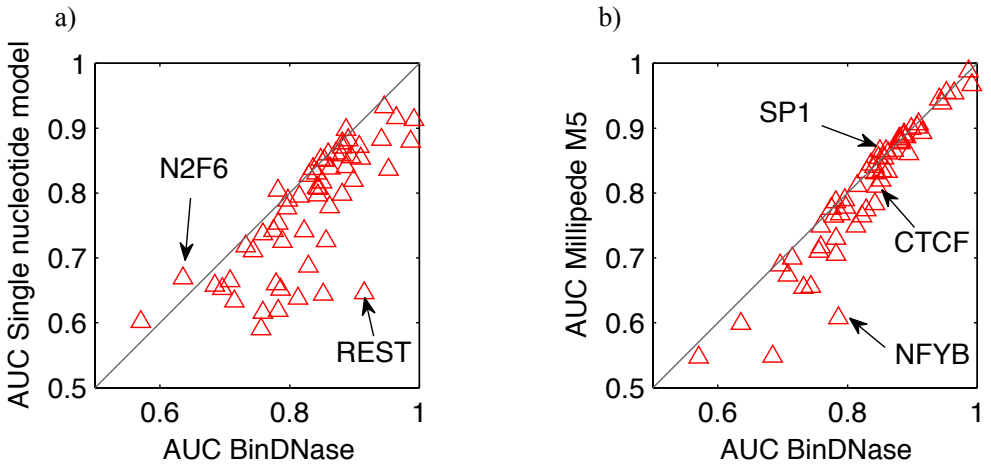


Figure 6.8: Feature selection improves TF binding predictions. a) The AUC scores for BinDNase and a modified version of BinDNase without feature selection and each bin at single nucleotide resolution. b) The AUC scores for BinDNase and MILLIPEDE. Negative sets 2 (the hotspots) were used in these comparisons.

tion 3.2, [7]). Here we demonstrate that these high resolution features can be used to improve binding predictions.

Comparison between MILLIPEDE and BinDNase for the 57 proteins gives expected results as the prediction performance is equal or better for all TFs in this study. BinDNase can be viewed as a more general version of the MILLIPEDE algorithm: instead of assigning bins similarly for each transcription factor our method finds optimal features (data binning) for each TF. In the worst case, BinDNase should find similar features and, thereby, similar prediction performance as MILLIPEDE, if those indeed happens to be optimal. For many proteins, BinDNase can improve the prediction accuracy as shown in Figure 6.8 b). The increase in performance is acquired by using higher resolution features and by assigning the bins differently for each transcription factor. BinDNase increases the prediction performance the most for TFs for which the average DNase I cleavage does not follow the canonical footprint pattern assumed in MILLIPEDE, such as NFYB (see Figure 6.6 d) for the average DNase signal and the trained model).

6.7 Prediction accuracy saturates at a modest sequencing depth

The effect of sequencing depth on prediction accuracy was investigated by making predictions on the candidate binding sites using only subsamples of all reads (Figure 6.9). For many TFs for which the predictions are easy to make (i.e., high AUC) the required sequencing depth is much lower than the depth in the ENCODE DNase-seq data sets. Prediction accuracy saturates already using 30M-60M of the 270M original DNase-seq reads. For TFs which are more difficult to predict (AUC 0.8 or below) sufficient saturation is achieved at between 150M-200M reads per sample. For a few TFs, such as EGR1 and RFX5, it seems that deeper sequencing could further improve the prediction results.

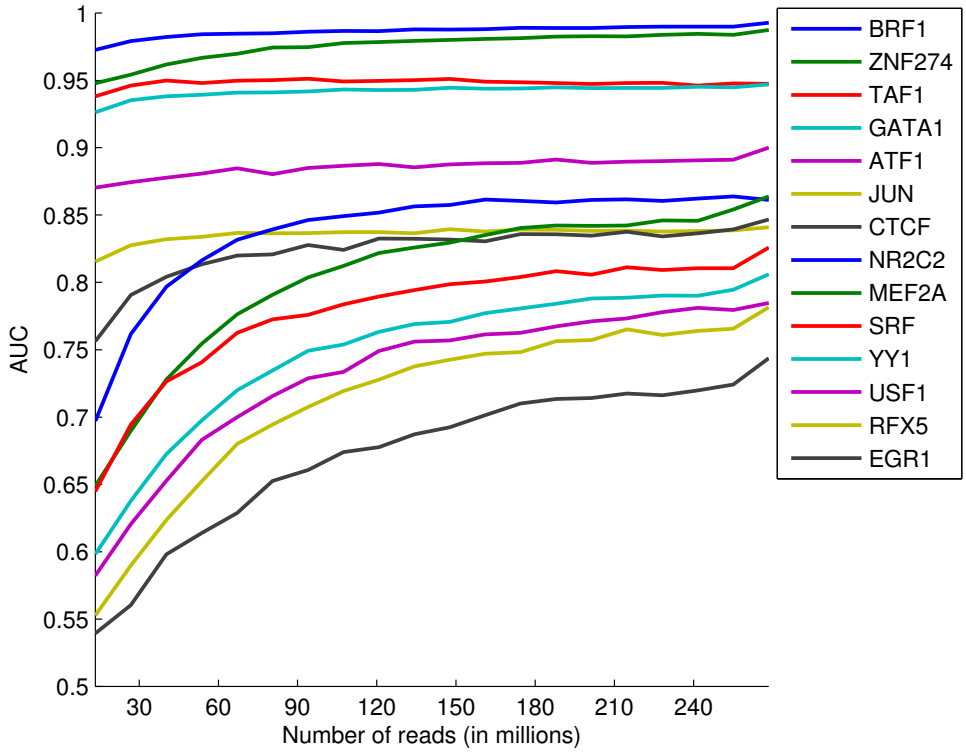


Figure 6.9: Effect of sequencing depth on TF binding prediction accuracy. The AUC scores are computed for a representative collection of TFs using subsampled versions of the original DNase-seq data. The original sequencing depth is 270M reads.

6.8 BinDNase generalizes between different cell-types

As the goal is to build universal transcription factor binding models, it is highly important that the models can produce useful predictions in different conditions and cell types. To test whether the models generalize for different cell types we used K562 and HepG2 for training the models and made the predictions for cell type K562, and vice versa. The models' prediction performances are very similar between cell types for all except a handful of TFs as shown in Figure 6.10. Notable differences between prediction accuracies between cell types are observed only for proteins with a low AUC score in both cell types. Naturally, whenever there is a difference in the model performance, the predictions made to the same cell type that the model was built with are more accurate.

Some of the proteins that behave similarly or differently between cell types are highlighted in Figure 6.10. Proteins that do not generalise well belong to e.g. helix-loop-helix (HLH) (BHLHE40) and leucine zipper families (JUND). USF2 contains both HLH and leucine zipper domains. Structural classes may be the reason for worse generalization capabilities. Models for proteins CTCF and TAF1 are highly similar no matter which cell type they are trained with and therefore the models produce highly similar prediction results.

Figure 6.11 emphasizes how the BinDNase models can be highly similar

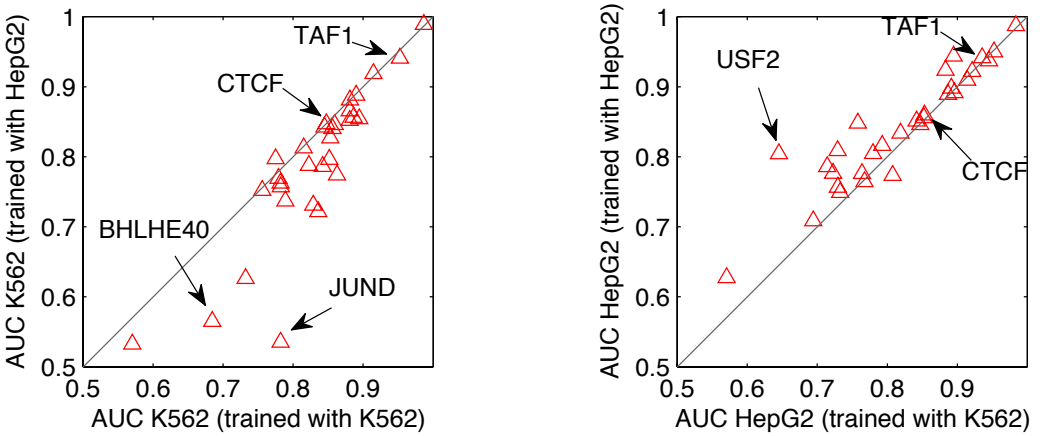


Figure 6.10: The models generalize well to different cell types. The model trained on another celltype works almost equally as well as the one trained with the same celltype that the predictions are made to. Negative sets 2 (the hotspots) were used in the modeling.

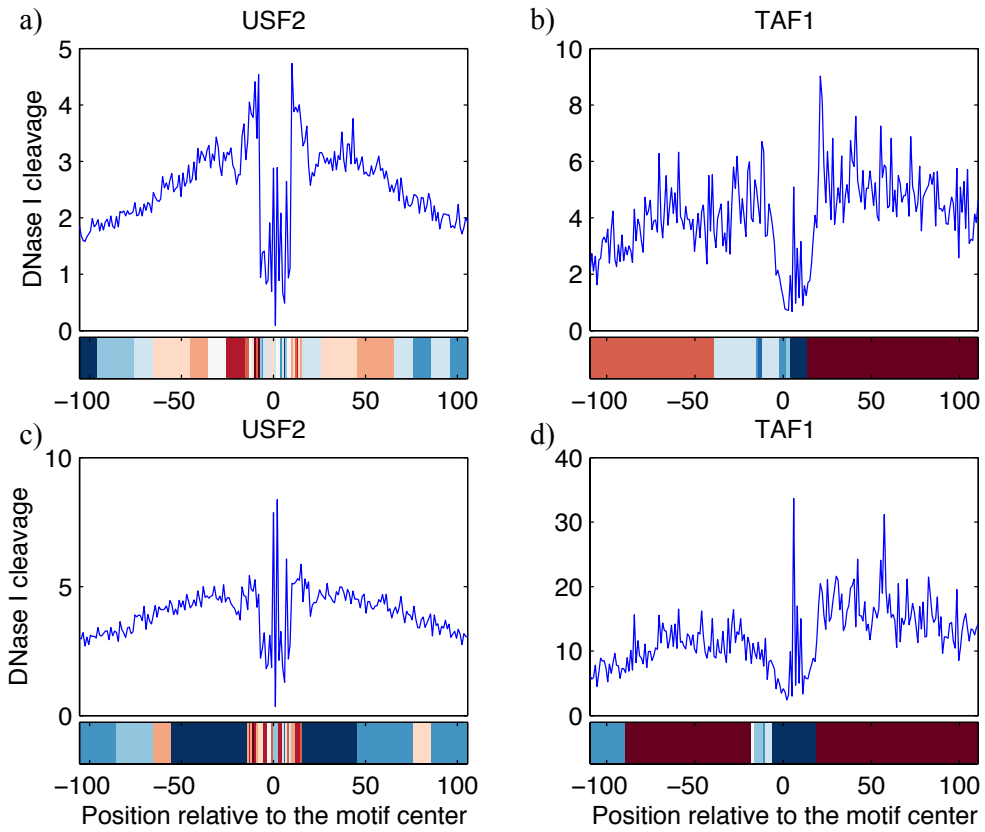


Figure 6.11: The upper panels show the average DNase I cleavage centered at the TF binding motifs. The coloured bars indicate the optimised feature selection and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient. The models trained with different cell types can be different (USF2) or highly similar (TAF1). a) USF2 (K562), b) TAF1 (K562), c) USF2 (HepG2), and d) TAF1 (HepG2).

for some TFs (TAF1), but slightly problematic for some TFs (USF2). The models trained with different cell types for TAF1 (Figures 6.11 b) and c)) are very similar as the models use reads falling to two wide bins flanking the binding site to increase the binding score (red bins) and reads falling to narrower bins in the candidate binding site to decrease the binding score (blue bins) for both cell types. There are no clear similarities in the models trained for USF2 which explains the differential binding prediction capabilities between the cell types. There are also clear differences in the average profiles between cell types as the DNase activity right in the middle of the candidate binding sites exceeds clearly the activity in the flanking regions for cell type HepG2 (Figure 6.11 c)).

Chapter 7

Discussion

Although transcription factor binding prediction using various high-throughput sequencing data has made significant progress recently, there is still an urgent demand to develop novel computational methods for analysing heterogeneous sequencing data sets. This work sheds light on many questions considering the computational analysis of deeply sequenced DNase I hypersensitivity data. Foremost, despite the fact that most previous methods have used traditional, lower resolution canonical DNase I hotspot or footprint models, we demonstrate that DNase-seq data contain high resolution information about TF-DNA interactions which can be used to improve discrimination between bound and unbound motif locations. The use of single nucleotide resolution DNase-seq data is hindered by the fact that not all signals in the data discriminate bound and unbound sites. Thus, efficient feature selection methods, such as the one used in this work, are needed to construct accurate TF binding prediction models. High resolution footprints are also TF-specific and hence the prediction models need to be constructed separately for each TF.

We developed a novel method, BinDNase, for TF binding prediction using DNase-seq data. Via comprehensive simulations we show that BinDNase performs better than existing methods. We show that BinDNase's prediction accuracy is generally well-saturated at the sequencing depths of the currently available DNase-seq data sets and that the method also generalises between cell types.

We also show in this work that even small discrepancies in DNase-seq data preprocessing will distort the nucleotide level footprint signal and consequently make the data considerable less informative about TF-specific binding sites. Our results demonstrate that the discriminatory model needs to be constructed separately for each TF, which reflects the fact that each TF has a specific TF-DNA interface and hence pose different DNase I cleavage

profiles at and around the exact motif location. We also show that extracting features from DNase-seq at multiple resolution, i.e. both nucleotide and lower resolution features, improves TF binding prediction accuracy and that sequencing coverage will have only a modest effect of the TF binding prediction accuracy. Taking BinDNase's high prediction performance and generalization capabilities into account, BinDNase is a versatile tool for accurate TF binding prediction. We believe that BinDNase will be a useful tool in practise and help revealing the mechanisms of transcriptional regulation in numerous applications.

7.1 Future research directions

In the future BinDNase will be used for studying variation in transcription factor binding among humans using DNase I hypersensitivity data. The main application will be detecting differences in transcription factor binding that lead to phenotypical effects. These detected phenotypical changes often contribute to human health and therefore detecting mutations affecting individual's phenotype via differential transcription factor binding is of high importance.

The most relevant part of the research to the real world applications is the identification of biologically relevant genomic sites that are differentially bound by TFs between individuals. These differences will be linked to phenotypical differences such as gene expression and disease. A method for distinguishing differential binding utilizing BinDNase predictions has to be developed in order to research this question. The ultimate goal is to find causal mechanisms behind different traits and to explain phenotypical phenomena caused by transcription factor binding.

Another prospective research line is to combine the BinDNase binding models of each TF into a multi-class classifier which predicts the binding of all TFs genome wide. This classifier could also take into account the competitive binding between TFs. It is also known that instead of acting always separately TFs often bind together. The combinatorial effects of TFs have been studied but the phenomenon is still poorly understood for the majority of TFs. In addition to modeling competitive binding, multi-class BinDNase classifier could take the combinatorial binding effects into account as well.

Bibliography

[1] http://2011.igem.org/team:dtu-denmark/background_srna, (21.05.2014).

[2] https://howardhughes.trinity.duke.edu/blogs/2011/06/24/zinc-fingers-plasmids-and-my-battle-with-e-coli, (21.05.2014).

[3] M. Annala, K. Laurila, H. Lähdesmäki, and M. Nykter. A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS ONE*, 6(5):e20059, 2011.

[4] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283 – 291, 2004.

[5] M. Berger, A. Philippakis, A. Qureshi, F. He, P. Estep III, and B. M.L. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, 2006.

[6] C. E. Grant, T. L. Bailey, and W. S. Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[7] H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, and M. Brown. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods*, 11:73–78, 2014.

[8] S. John, P. J. Sabo, R. E. Thurman, M.-H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43:264–268, 2011.

[9] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. Vaquerizas, J. Yan, M. Sillanpää, M. Bonke, K. Palin,

S. Talukder, T. Hughes, N. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, 2010.

[10] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. Vaquerizas, R. Vincentelli, N. Luscombe, T. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[11] J. Kähärä and H. Lähdesmäki. Evaluating a linear k-mer model for protein-dna interactions using high-throughput selex data. *BMC Bioinformatics*, 14(Suppl 10), 2013.

[12] M. Karin, Z. gang Liu, and E. Zandi. Ap-1 function and regulation. *Current Opinion in Cell Biology*, 9(2):240 – 246, 1997.

[13] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M.-Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder. Variation in transcription factor binding among humans. *Science*, 328(5975):232–235, 2010.

[14] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22(9):1813–1831, 2012.

[15] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[16] K. Laurila and H. Lähdesmäki. Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding. *In Silico Biology*, 9(4)(1386-6338):209–24, 2009.

[17] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[18] K. Luo and A. J. Hartemink. using dnase digestion data to accuractely identify transcription factor binding sites. In *Biocomputing*, pages 80–91, 2013.

[19] P. Madrigal and P. Krajewski. Current bioinformatic approaches to identify DNase i hypersensitive sites and genomic footprints from DNase-seq data. *Frontiers in Genetics*, 3:PMC3484326, 2012.

[20] E. R. Mardis. Chip-seq: welcome to the new frontier. *Nat Meth*, 4(8):613–614, Aug. 2007.

[21] M. Matsuda, N. Sakamoto, and Y. Fukumaki. δ-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the δ-globin gene promoter. *Blood*, 80:1347–1351, 1992.

[22] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, and A. K. Johnson. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

[23] C. O. Pabo and R. T. Sauer. Protein-dna recognition. *Annual Review of Biochemistry*, 53(1):293–321, 1984. PMID: 6236744.

[24] J. Piper, M. C. Elze, P. Cauchy, P. N. Cockerill, C. Bonifer, and S. Ott. Wellington: a novel method for the accurate identification of digital genomic footprints from dnase-seq data. *Nucleic Acids Research*, 41(21):e201, 2013.

[25] R. Pique-Regi, J. F. Degner, and A. A. P. et al. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Research*, 21:447–455, 2011.

[26] K. Samejima and W. C. Earnshaw. Trashing the genome: the role of nucleases during apoptosis. *Nature Reviews Molecular Cell Biology*, 6:677–688, 2005.

[27] R. Staden. Staden: Searching for motifs in nucleic acid sequences. In *Computer Analysis of Sequence Data*, volume 25 of *Methods in Molecular Biology*, pages 93–102. Springer New York, 1994.

[28] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2012.

[29] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263, Apr. 2009.

[30] J. Wang, J. Zhuang, S. lyer, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22:1798–1812, 2012.

[31] M. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. Bussemaker, M. Quaid, M. Bulyk, G. Stolovitzky, T. Hughes, P. Agius, A. Arvey, P. Bucher, C. Callan Jr., C. Chang, C.-Y. Chen, Y.-S. Chen, Y.-W. Chu, J. Grau, I. Grosse, V. Jagannathan, J. Keilwagen, S. Kiebasa, J. Kinney, H. Klein, M. Kursa, H. Lähdesmäki, K. Laurila, C. Lei, C. Leslie, C. Linhart, A. Murugan, A. MysickovÃ¡, W. Noble, M. Nykter, Y. Orenstein, S. Posch, J. Ruan, W. Rudnicki, C. Schmid, R. Shamir, W.-K. Sung, M. Vingron, and Z. Zhang. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, 2013.

[32] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer. Transfac: an integrated system for gene expression regulation. *Nucleic Acids Research*, 28(1):316–319, 2000.

[33] E. Wingender, T. Schoeps, and J. Dönitz. Tfclass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, 41(D1):D165–D170, 2012.

[34] J. Wittwer, J. Marti-Jaun, and M. Hersberger. Functional polymorphism in ALOX15 results in increased allele-spesific transcription in macrophages through binding of the transcription factor SPI1. *Human Mutation*, 27:78–87, 2006.

Appendix A

ChIP-seq datasets

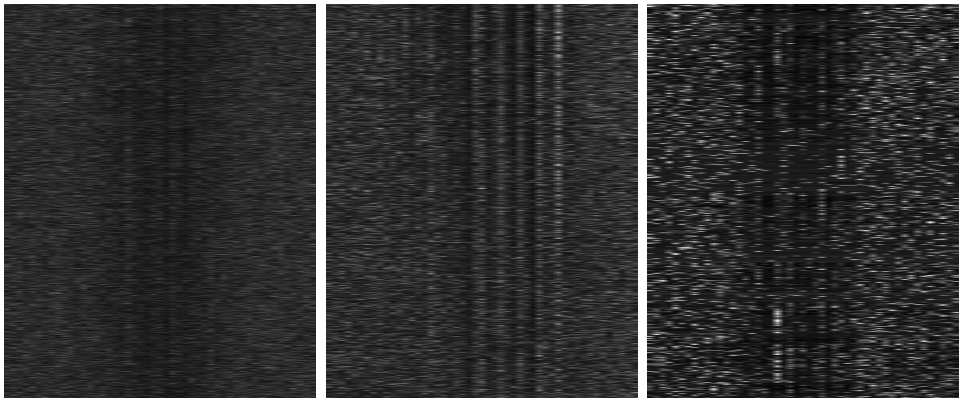
TF	Celltype	Filename
BACH1	K562	wgEncodeAwgTfbsSydhK562Bach1sc14700IggrabUniPk.narrowPeak
GATA1	K562	wgEncodeAwgTfbsSydhK562Gata1UcdUniPk.narrowPeak
MAFK	K562	wgEncodeAwgTfbsSydhK562Mafkab50322IggrabUniPk.narrowPeak
CEBPB	K562	wgEncodeAwgTfbsSydhK562CebpbIggrabUniPk.narrowPeak
E2F4	K562	wgEncodeAwgTfbsSydhK562E2f4UcdUniPk.narrowPeak
EGR1	K562	wgEncodeAwgTfbsHaibK562Egr1V0416101UniPk.narrowPeak
ELF1	K562	wgEncodeAwgTfbsHaibK562Elf1sc631V0416102UniPk.narrowPeak
ELK1	K562	wgEncodeAwgTfbsSydhK562Elk112771IggrabUniPk.narrowPeak
CTCFL	K562	wgEncodeAwgTfbsHaibK562Ctcflsc98982V0416101UniPk.narrowPeak
FOSL1	K562	wgEncodeAwgTfbsHaibK562Fosl1sc183V0416101UniPk.narrowPeak
FOS	K562	wgEncodeAwgTfbsSydhK562CfosUniPk.narrowPeak
MAFF	K562	wgEncodeAwgTfbsSydhK562MaffIggrabUniPk.narrowPeak
MXI1	K562	wgEncodeAwgTfbsSydhK562Mxi1af4185IggrabUniPk.narrowPeak
RFX5	K562	wgEncodeAwgTfbsSydhK562Rfx5IggrabUniPk.narrowPeak
SMC3	K562	wgEncodeAwgTfbsSydhK562Smc3ab9263IggrabUniPk.narrowPeak
RAD21	K562	wgEncodeAwgTfbsHaibK562Rad21V0416102UniPk.narrowPeak
STAT5	K562	wgEncodeAwgTfbsHaibK562Stat5asc74442V0422111UniPk.narrowPeak
SP2	K562	wgEncodeAwgTfbsHaibK562Sp2sc643V0416102UniPk.narrowPeak
TAF1	K562	wgEncodeAwgTfbsHaibK562Taf1V0416101UniPk.narrowPeak
USF2	K562	wgEncodeAwgTfbsSydhK562Usf2IggrabUniPk.narrowPeak
ZBTB33	K562	wgEncodeAwgTfbsHaibK562Zbtb33Pcr1xUniPk.narrowPeak
ZBTB7A	K562	wgEncodeAwgTfbsHaibK562Zbtb7asc34508V0416101UniPk.narrowPeak
ZNF143	K562	wgEncodeAwgTfbsSydhK562Znf143IggrabUniPk.narrowPeak
ATF3	K562	wgEncodeAwgTfbsHaibK562Atf3V0416101UniPk.narrowPeak
BDP1	K562	wgEncodeAwgTfbsSydhK562Bdp1UniPk.narrowPeak
BHLHE40	K562	wgEncodeAwgTfbsSydhK562Bhlhe40nb100IggrabUniPk.narrowPeak
GABPA	K562	wgEncodeAwgTfbsHaibK562GabpV0416101UniPk.narrowPeak
CTCF	K562	wgEncodeAwgTfbsBroadK562CtcfUniPk.narrowPeak
JUND	K562	wgEncodeAwgTfbsSydhK562JundIggrabUniPk.narrowPeak
NR2C2	K562	wgEncodeAwgTfbsSydhK562Tr4UcdUniPk.narrowPeak
NR2F2	K562	wgEncodeAwgTfbsHaibK562Nr2f2sc271940V0422111UniPk.narrowPeak
E2F6	K562	wgEncodeAwgTfbsHaibK562E2f6V0416102UniPk.narrowPeak
ETS1	K562	wgEncodeAwgTfbsHaibK562Ets1V0416101UniPk.narrowPeak
SP1	K562	wgEncodeAwgTfbsHaibK562Sp1Pcr1xUniPk.narrowPeak
USF1	K562	wgEncodeAwgTfbsHaibK562Usf1V0416101UniPk.narrowPeak
JUN	K562	wgEncodeAwgTfbsSydhK562CjunUniPk.narrowPeak
THAP1	K562	wgEncodeAwgTfbsHaibK562Thap1sc98174V0416101UniPk.narrowPeak
JUNB	K562	wgEncodeAwgTfbsUchicagoK562EjunbUniPk.narrowPeak
MAX	K562	wgEncodeAwgTfbsHaibK562MaxV0416102UniPk.narrowPeak
MEF2A	K562	wgEncodeAwgTfbsHaibK562Mef2aV0416101UniPk.narrowPeak
MYC	K562	wgEncodeAwgTfbsSydhK562CmycIggrabUniPk.narrowPeak

NFE2	K562	wgEncodeAwgTfbsSydhK562Nfe2UniPk.narrowPeak
REST	K562	wgEncodeAwgTfbsHaibK562NrsfV0416102UniPk.narrowPeak
NFYA	K562	wgEncodeAwgTfbsSydhK562NfyaUniPk.narrowPeak
NFYB	K562	wgEncodeAwgTfbsSydhK562NfybUniPk.narrowPeak
EP300	K562	wgEncodeAwgTfbsSydhK562P300IggrabUniPk.narrowPeak
SPI1	K562	wgEncodeAwgTfbsHaibK562Pu1Pcr1xUniPk.narrowPeak
SRF	K562	wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak
TBP	K562	wgEncodeAwgTfbsSydhK562TbpIggmusUniPk.narrowPeak
TEAD4	K562	wgEncodeAwgTfbsHaibK562Tead4sc101184V0422111UniPk.narrowPeak
NRF1	K562	wgEncodeAwgTfbsSydhK562Nrf1IggrabUniPk.narrowPeak
YY1	K562	wgEncodeAwgTfbsHaibK562Yy1V0416102UniPk.narrowPeak
BRF1	K562	wgEncodeAwgTfbsSydhK562Brf1UniPk.narrowPeak
ZNF263	K562	wgEncodeAwgTfbsSydhK562Znf263UcdUniPk.narrowPeak
ATF1	K562	wgEncodeAwgTfbsSydhK562Atf106325UniPk.narrowPeak
ZNF274	K562	wgEncodeAwgTfbsSydhK562Znf274m01UcdUniPk.narrowPeak
GATA2	K562	wgEncodeAwgTfbsSydhK562Gata2UcdUniPk.narrowPeak

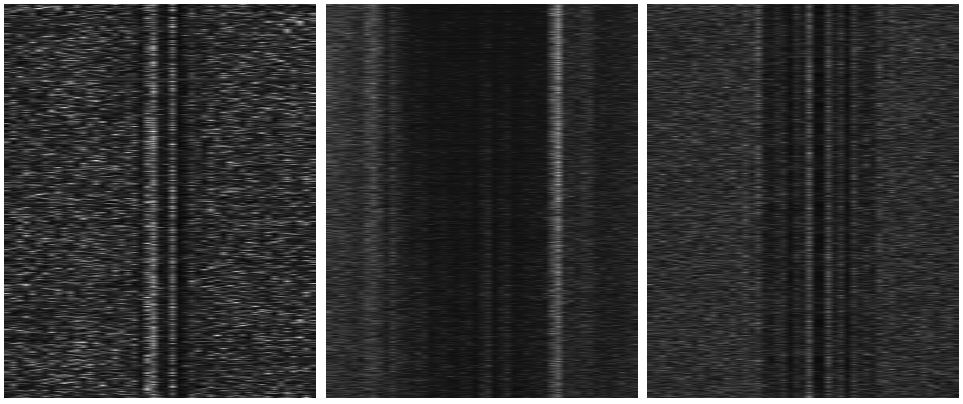
TF	Celltype	Filename
JUN	HepG2	wgEncodeAwgTfbsSydhHepg2CjunIggrabUniPk.narrowPeak
BHLHE40	HepG2	wgEncodeAwgTfbsSydhHepg2Bhlhe40cIggrabUniPk.narrowPeak
CEBPB	HepG2	wgEncodeAwgTfbsSydhHepg2CebpbIggrabUniPk.narrowPeak
ELF1	HepG2	wgEncodeAwgTfbsHaibHepg2Elf1sc631V0416101UniPk.narrowPeak
GABPA	HepG2	wgEncodeAwgTfbsHaibHepg2GabpPcr2xUniPk.narrowPeak
JUND	HepG2	wgEncodeAwgTfbsSydhHepg2JundIggrabUniPk.narrowPeak
MAFF	HepG2	wgEncodeAwgTfbsSydhHepg2Maffm8194IggrabUniPk.narrowPeak
SP2	HepG2	wgEncodeAwgTfbsHaibHepg2Sp2V0422111UniPk.narrowPeak
TAF1	HepG2	wgEncodeAwgTfbsHaibHepg2Taf1Pcr2xUniPk.narrowPeak
USF1	HepG2	wgEncodeAwgTfbsHaibHepg2Usf1Pcr1xUniPk.narrowPeak
YY1	HepG2	wgEncodeAwgTfbsHaibHepg2Yy1sc281V0416101UniPk.narrowPeak
SRF	HepG2	wgEncodeAwgTfbsHaibHepg2SrfV0416101UniPk.narrowPeak
EP300	HepG2	wgEncodeAwgTfbsHaibHepg2P300V0416101UniPk.narrowPeak
CTCF	HepG2	wgEncodeAwgTfbsHaibHepg2Ctcfsc5916V0416101UniPk.narrowPeak
ZNF274	HepG2	wgEncodeAwgTfbsSydhHepg2Znf274UcdUniPk.narrowPeak
ATF3	HepG2	wgEncodeAwgTfbsHaibHepg2Atf3V0416101UniPk.narrowPeak
MAFK	HepG2	wgEncodeAwgTfbsSydhHepg2Mafkab50322IggrabUniPk.narrowPeak
RFX5	HepG2	wgEncodeAwgTfbsSydhHepg2Rfx5200401194IggrabUniPk.narrowPeak
SP1	HepG2	wgEncodeAwgTfbsHaibHepg2Sp1Pcr1xUniPk.narrowPeak
USF2	HepG2	wgEncodeAwgTfbsSydhHepg2Usf2IggrabUniPk.narrowPeak
MXI1	HepG2	wgEncodeAwgTfbsSydhHepg2Mxi1UniPk.narrowPeak
MYC	HepG2	wgEncodeAwgTfbsUtaHepg2CmycUniPk.narrowPeak
RAD21	HepG2	wgEncodeAwgTfbsHaibHepg2Rad21V0416101UniPk.narrowPeak
REST	HepG2	wgEncodeAwgTfbsHaibHepg2NrsfV0416101UniPk.narrowPeak
MAX	HepG2	wgEncodeAwgTfbsSydhHepg2MaxIggrabUniPk.narrowPeak
SMC3	HepG2	wgEncodeAwgTfbsSydhHepg2Smc3ab9263IggrabUniPk.narrowPeak
NRF1	HepG2	wgEncodeAwgTfbsSydhHepg2Nrf1IggrabUniPk.narrowPeak
ZBTB7A	HepG2	wgEncodeAwgTfbsHaibHepg2Zbtb7aV0416101UniPk.narrowPeak
TBP	HepG2	wgEncodeAwgTfbsSydhHepg2TbpIggrabUniPk.narrowPeak
TEAD4	HepG2	wgEncodeAwgTfbsHaibHepg2Tead4sc101184V0422111UniPk.narrowPeak
NR2C2	HepG2	wgEncodeAwgTfbsSydhHepg2Tr4UcdUniPk.narrowPeak

Appendix B

Heatmaps

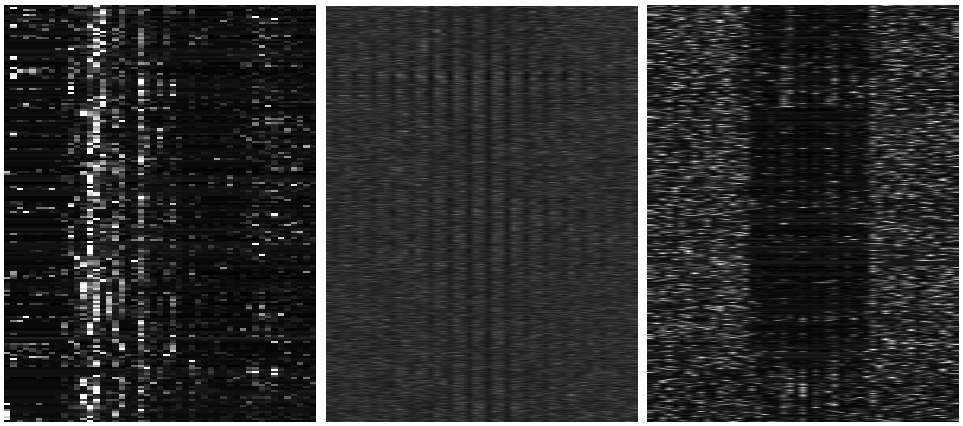


(a) ZBTB7A (b) ZNF143 (c) ATF3

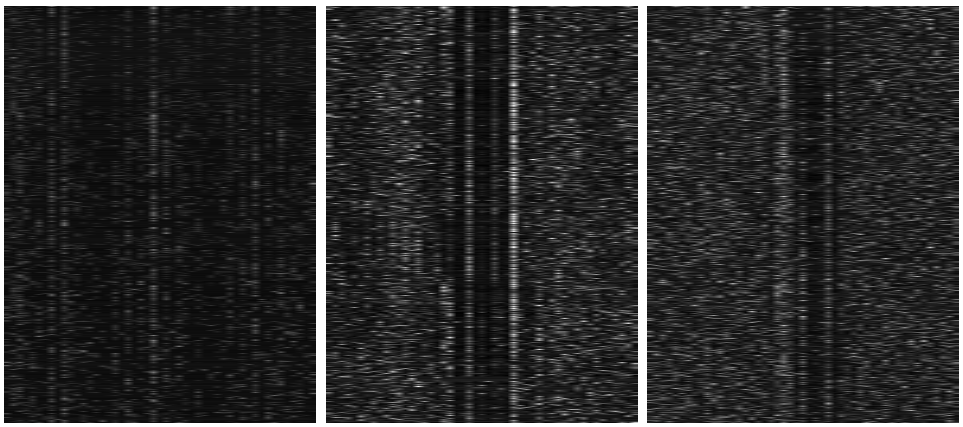


(d) ELK1 (e) CTCFL (f) FOSL1

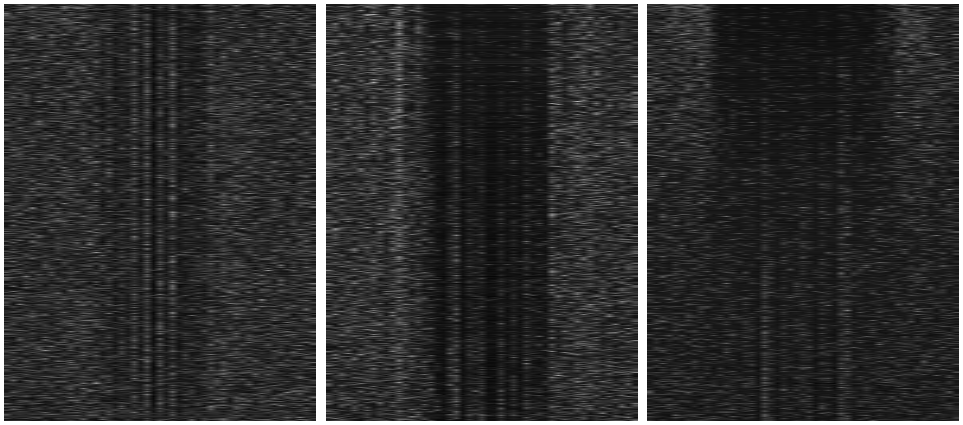
Figure B.1: Heatmaps showing the data at the candidate binding sites for different proteins.



(a) BRF1 (b) ZNF263 (c) ATF1

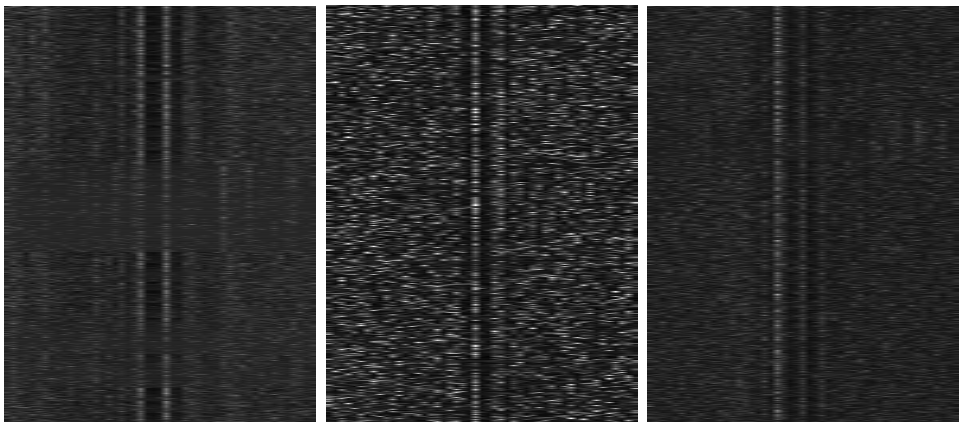


(d) ZNF274 (e) GATA2 (f) MEF2A

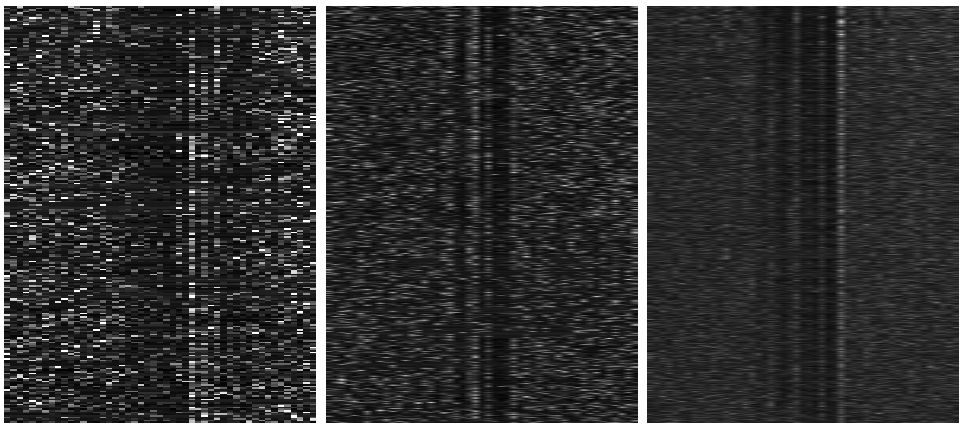


(g) MYC (h) NFE2 (i) REST

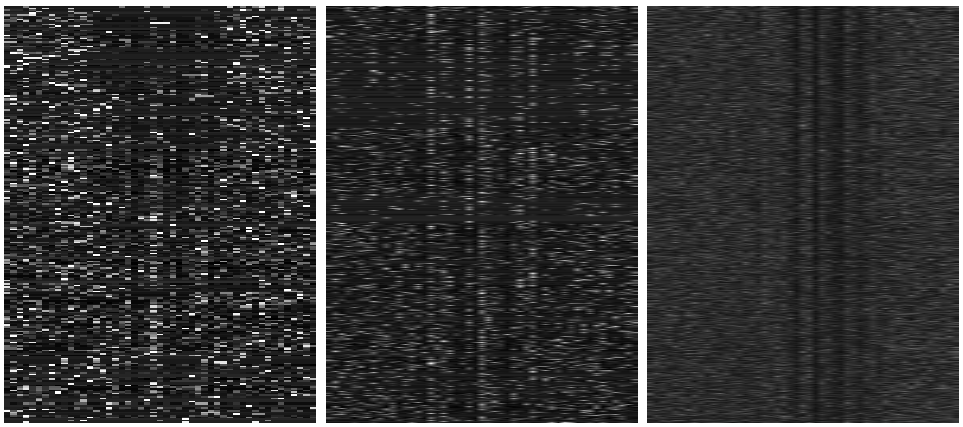
Figure B.2: Heatmaps showing the data at the candidate binding sites for different proteins.



(a) NFYB (b) EP300 (c) SPI1

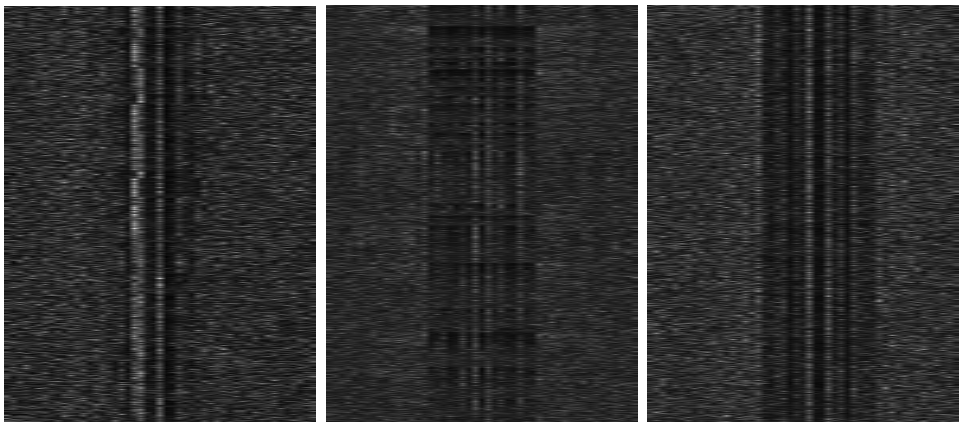


(d) TBP (e) TEAD4 (f) YY1

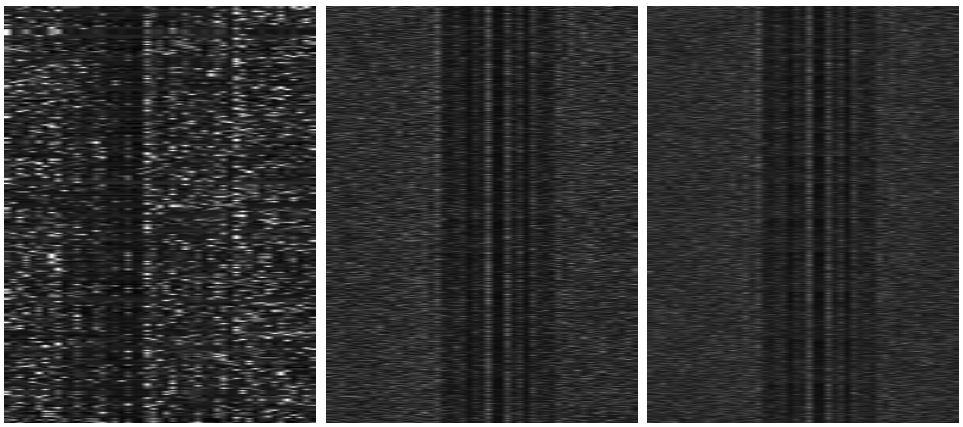


(g) NR2C2 (h) NR2F2 (i) E2F6

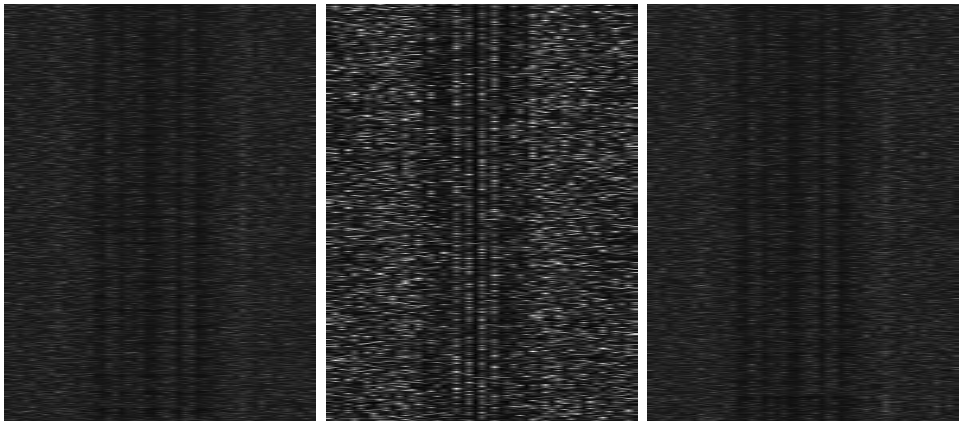
Figure B.3: Heatmaps showing the data at the candidate binding sites for different proteins.



(a) GABPA (b) USF1 (c) JUN

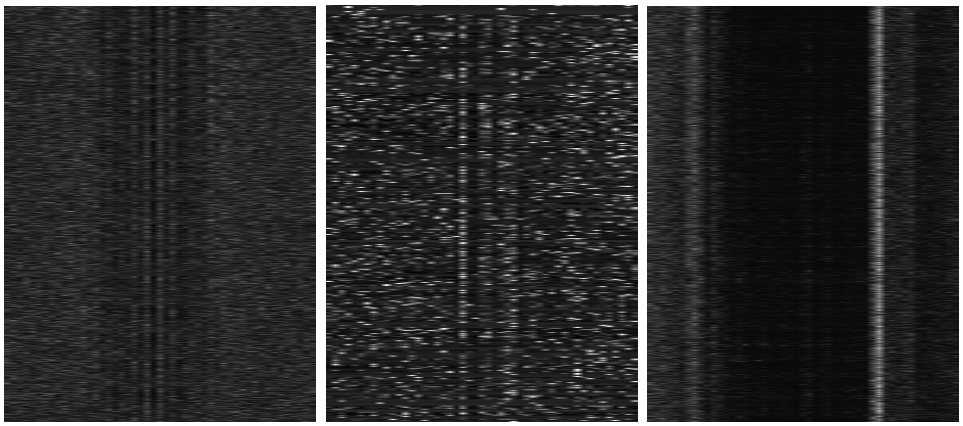


(d) THAP1 (e) JUNB (f) JUND

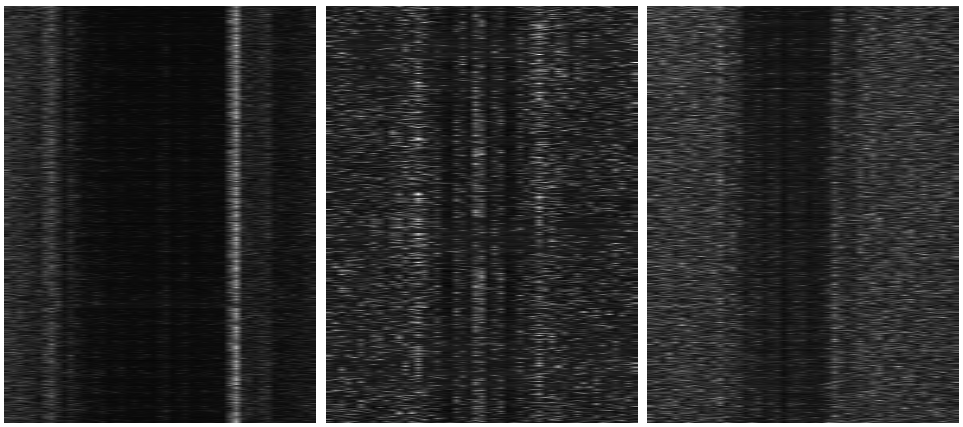


(g) MAFF (h) MXI1 (i) MAFK

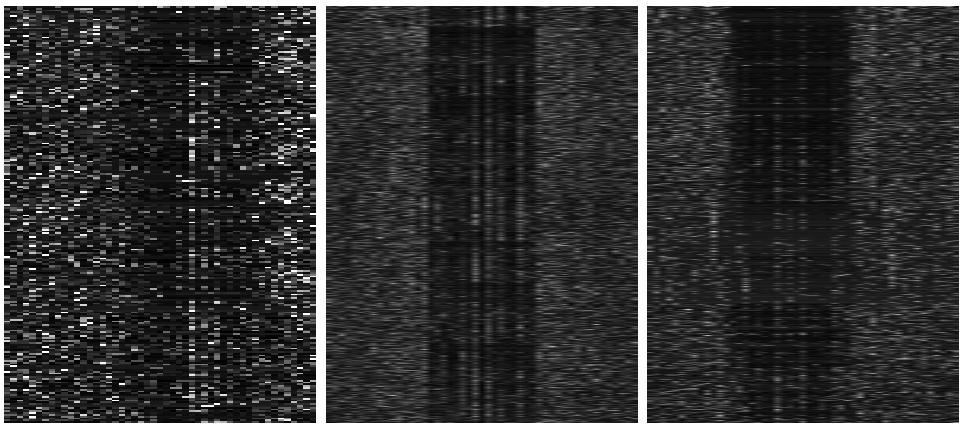
Figure B.4: Heatmaps showing the data at the candidate binding sites for different proteins.



(a) MAX (b) RFX5 (c) SMC3

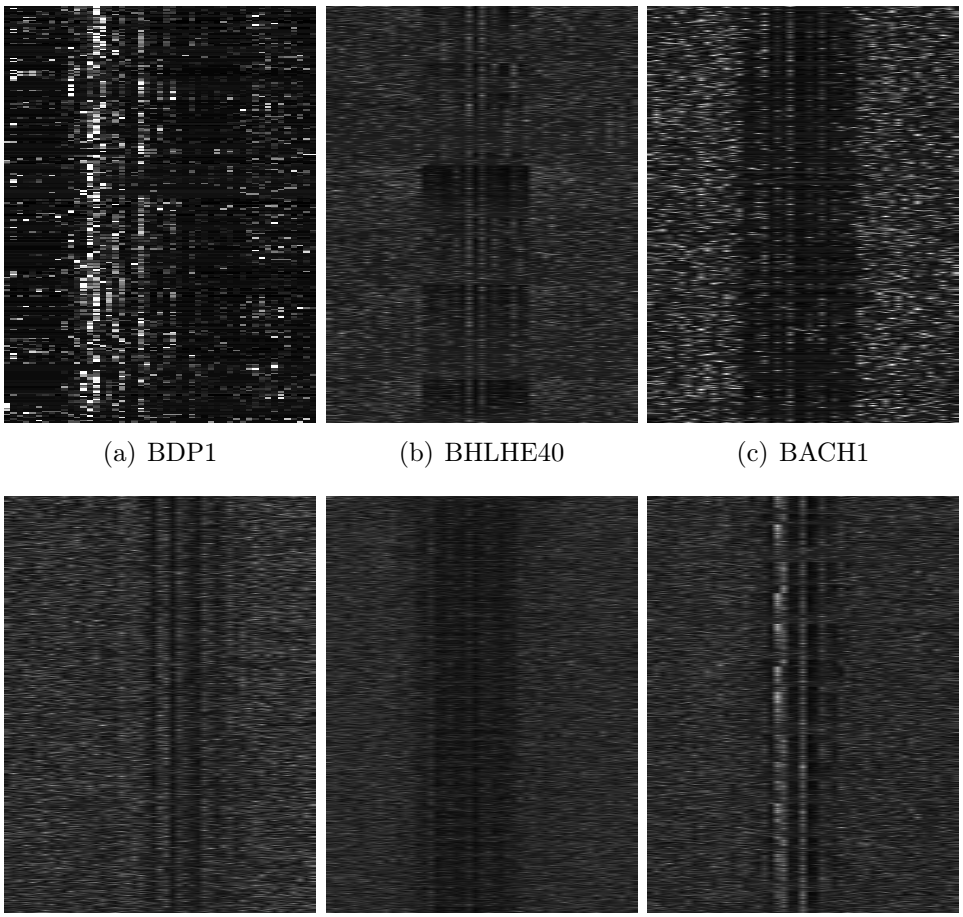


(d) RAD21 (e) STAT5 (f) SP2



(g) TAF1 (h) USF2 (i) ZBTB33

Figure B.5: Heatmaps showing the data at the candidate binding sites for different proteins.



(a) BDP1 (b) BHLHE40 (c) BACH1

(d) E2F4 (e) EGR1 (f) ELF1

Figure B.6: Heatmaps showing the data at the candidate binding sites for different proteins.